



CYBERHATE AND HUMAN RIGHTS

DR EMMA A. JANE
SENIOR LECTURER
SCHOOL OF THE ARTS AND MEDIA
UNSW SYDNEY



DR NICOLE A VINCENT
SENIOR LECTURER, AND xFUTURES LAB C.I.
FACULTY OF TRANSDISCIPLINARY INNOVATION
UNIVERSITY OF TECHNOLOGY SYDNEY



SECTION TITLE	PAGE
1. LANGUAGE WARNING	3
2. ABOUT THIS DOCUMENT	4
2.1. Structure	4
2.2. Sources	6
2.3. Authors	18
2.4. Disclaimers	20
2.5. Limitations	21
3. CYBERHATE	22
3.1. The rise of cyberhate	24
3.2. Gendered characteristics	25
3.3. Impacts and harms	28
3.4. Workplace harassment	29
3.5. New digital divide	34
4. RESPONSES TO CONSULTATION QUESTIONS	36
4.1. Response to Question One	37
4.2. Response to Question Two	46
4.3. Response to Question Three	58
4.4. Response to Question Four	66
4.5. Response to Question Five	68
5. REFERENCES	69

1. LANGUAGE WARNING

This report includes examples of real-life gendered cyberhate, some of which involves explicit imagery of sexual violence, as well as extreme racism and homophobia. Many readers are likely to find it confronting and offensive. While our intention is not to cause anyone gratuitous upset, in our view citing unexpurgated examples of cyberhate is essential if we wish to convey the nature and force of contemporary misogyny online. This is because – as we go on to explain in Section 4.2.1. of this submission – the use of generic descriptors such as “hostile”, “graphic”, and “in bad taste” simply do not capture the threatening nature and violence of the phenomenon in such a way that it can even be properly understood and conceptualised, let alone addressed.

2. ABOUT THIS DOCUMENT

In this section we do five things. In Section 2.1. we provide a brief overview of how this submission has been structured. In Section 2.2. we discuss the sources we have drawn on in putting together this submission — namely, Dr Emma A. Jane’s work on the topic of cyberhate funded by the Australian Research Council which critically informs our approach to the topic of cyberhate, key insights gained from the Cyberhate Symposium that inform our answers to the AHRC consultation questions, and our prior published work which too has influenced this submission’s shape and content. In Section 2.3. we provide background information about the authors of this submission and their fields of research expertise. In Section 2.4. we note that the views expressed in this submission are our own and not necessarily those of our respective employers, affiliated institutions, or research funding bodies. Finally, in Section 2.5. we explain why this submission offers responses only to the first five consultation questions listed in the Australian Human Rights Commission’s (AHRC’s) Human Rights and Technology (HRT) Issues Paper [58], but not to the last five questions.

2.1. STRUCTURE

This section provides a brief overview of how this document is structured. That is, it aims to not only sketch what information is contained in each of the sections of this submission, but also to explain how the different sections relate to one another. Given the AHRC’s interest in the ten consultation questions that were listed in the HRT Issues Paper, for convenience we shall start by discussing the section that contains our responses to those questions, and then we shall work our way backwards from there.¹

Our answers to the AHRC’s first five consultation questions² are contained in Section 4 of this submission. Our response to question 1 draws on examples discussed throughout this submission to demonstrate how the phenomenon we call “cyberhate”, which is critically reliant on a range of technologies associated with the internet, infringes on at least ten of the Articles defined in the Universal Declaration of Human Rights (UDHR). Our response to question 2

¹ In this discussion we sometimes refer to the Simplified Version of the full text of the UDHR [1], but on other occasions we refer to the original wording of the UDHR [2]. We switch between references to these two documents mainly because sometimes the wording of one document is better-attuned to help us convey the points which we are making.

² We explain why we only address those but not the remaining questions in Section 2.5. below.

draws on a critically important observation from the Cyberhate Symposium³ that we convened last year, and on which we officially report for the very first time in the present submission, to highlight a range of insidious conceptual, structural, institutional, and technological factors – as well as interactions between them – that make it extremely difficult to tackle cyberhate without summoning the support of the AHRC. Given the issues that we highlight in our response to question 2, not surprisingly our responses to questions 3 and 4 – which again draw on observations from the Cyberhate Symposium – discuss how those issues can be addressed through legislative, regulative, design, and other responses. Specifically, our response to question 3 discusses how Australian law should protect human rights from cyberhate, focusing in particular on the need for legislative and regulative responses that encourage and mandate the teaching and use of Value-Sensitive Design, Default Choice Architectures, and Socially Responsible Innovation methodologies in the process of designing new technologies. By comparison, our response to question 4 discusses how the Australian Government, private sector, and others, can protect human rights from cyberhate, focusing in particular on technology manufacturers, the education sector, researchers, the media, police, activist groups, and stakeholders, as well as citing demonstrative examples of the kinds of technology that we believe could help. Finally, our brief response to question 5 discusses the potential for Artificial Intelligence techniques to create opportunities to protect human rights from cyberhate.

The discussion in the HRT Issues Paper clearly recognises that cyber abuse poses important human rights concerns. At the same time, though, having investigated this topic extensively for many years, we are also acutely aware of the fact that others do not share this view. The importance of cyberhate is often either played down or even completely dismissed out of hand by those claiming that cyberhate is not really such a serious problem since it is, after all, “just words” and/or “just the internet”. Given the AHRC’s intention to draw upon the submissions made in response to the HRT Issues Paper to formulate its own set of recommendations for how to protect human rights from (and how to protect them via) new technologies, we decided to also include a discussion in Section 3 of this submission that sets the record straight, and pre-emptively responds to those who might doubt the critical importance of cyberhate as a human rights issue. By doing this, we hope that those to whom the AHRC will make its own further recommendations will also recognise cyberhate as a critically important threat to human

³ We discuss the Cyberhate Symposium, and note its many distinguished participants to whom we are deeply grateful for contributing, in Section 2.2.2. below.

rights, and that they too will acknowledge that this threat stems from a constantly evolving new technology (the internet) that unfortunately keeps evolving in ways that, if left unaddressed, will continue to leave people vulnerable to exposure to cyberhate.

Given the large number of sources that we have drawn upon in formulating this submission, and given that some of those sources are (i) brand new and (ii) involved many people including members of parliament giving up their valuable time to participate in the above-mentioned Cyberhate Symposium, we use Sections 2.2. and 2.3. to acknowledge our sources and give credit where it is due, as well as to state our credentials. Furthermore, since this submission is the work of two academics with different institutional affiliations, in Section 2.4. we clearly state that the views expressed herein are not necessarily representative of the views of our respective employers or affiliated institutions, but our own views. Finally, as we noted in a footnote above, in Section 2.5. we explain why we have only provided responses to the AHRC's first five consultative questions but not to the rest of those questions.

Lastly, the references we have cited in this submission are contained in Section 5.

2.2. SOURCES

In compiling this submission, we have drawn upon three main sources other than the literature cited in Section 5, which we discuss in the next three sub-sections. In Section 2.2.1. we discuss Dr Emma A. Jane's work on the topic of cyberhate, since her expertise in this field clearly informs how we understand – and how we have approached the topic of cyberhate and the issues that it raises. In Section 2.2.2. we discuss the Cyberhate Symposium that was convened in 2017, and how it informed our responses to the AHRC's consultation questions. Finally, in Section 2.2.3. we also note our own prior published works, which have also informed this submission's shape and content.

2.2.1. *Cyberhate: the new digital divide?*

From January 2015 to December 2017, Dr Jane's research was funded by the Australian Research Council (ARC) in the form of a Discovery Early Career Research Award (DECRA). This was for a project called "Cyberhate: the new digital divide?" and involved studying rape threats and other hostility directed at women on the internet and on social media platforms. The aim was to investigate whether gendered cyberhate affected women's online participation, and whether it might constitute a new dimension of the digital divide. This study involved two main components: (i) mapping the history of misogyny online; and (ii) conducting qualitative

interviews with 52 Australian targets of gendered cyberhate conducted from 2015 through to 2017. It was approved by the UNSW Sydney Human Research Ethics Committee.⁴

Two main selection criteria were used to recruit interviewees: participants had to identify as female and to have experienced gendered cyberhate. A range of recruitment techniques were employed (for example, posters up at universities, social media, mail-outs, direct approaches to known targets, and purposive chain referral sampling). Two groups of interviewees were recruited to enable an exploration and comparison of the experiences of women in the public eye as well as “ordinary” women. While recruitment techniques were not designed to obtain a representative population sample, steps were taken to ensure that the interview cohort included women of colour, queer women, and Muslim women, as well as women from a range of age groups and socioeconomic (SES) circumstances.

The first group of interviewees (n=32) comprised women with public profiles who had experienced hostility or threats online, and had previously discussed this in public fora. These women had the option of being interviewed in an identifiable way using their real names, and most made use of this option. The sampling strategy for this group was purposive, aiming to target women from various demographics, locations, and socioeconomic (SES) groups who were regular users of the fora under inquiry and who had spoken about their previous cyberhate experiences in the media and/or in public domains. The rationale was that this group of subjects could potentially provide broad insights into the experience of receiving gendered cyberhate, as well as the ramifications of having spoken publicly about being a cyberhate target. The second group of interviewees (n=19) comprised women who were not (at the time they were interviewed) in public life, and who had experienced hostility or threats online but had not (at the time they were interviewed) spoken about these experiences in public fora. These interviewees all used pseudonyms and all identifying details were removed from their transcripts. Potential members of this group were excluded if they had spoken previously about their experiences of gendered cyberhate publicly or in the media. This was to enable the examination of episodes of gendered cyberhate that had not previously been reported by media outlets. In addition to providing a rich source of new data, this approach enabled the comparison of previously untold stories with those accounts currently circulating in the public domain. Dr Jane interviewed these women – aged between 19 to 52 – from 2015 through 2017.

⁴ The UNSW Sydney Research Ethics Committee reference for this project was HC15012.

Interviews were conducted in person or via Skype, with some follow-up interviews involving phone conversations and email.

All interviewees were invited to the Cyberhate Symposium, which we discuss in the next sub-section, and many of them attended.

Finally, this research project resulted in a wide variety of traditional and non-traditional scholarly research outputs, numerous media appearances, as well as other formal submissions, and unsurprisingly it has had a profound impact on the form and content of this submission.

2.2.2. Cyberhate Symposium

Secondly, some of the arguments and data that are found in this submission were obtained from a Cyberhate Symposium that the authors of this submission staged in Sydney in 2017, that subsequently received recognition from the NSW parliament in its unanimous passing of a motion congratulating Dr Jane and Dr Vincent for staging the event [20]. The present submission is the first occasion on which we officially report our full set of findings from this Symposium.

In addition to a large number of gendered cyberhate targets, participants at the Symposium included: politicians from different parties; representatives from various sectors of the police force; representatives from the Office of the eSafety Commissioner and other government agencies; professionals working in areas such as mental health, domestic violence, and education; people working in senior levels of IT management in corporate firms; platform managers and moderators; media personalities and journalists; coders, tech designers, and gamers; activists from various groups (including groups whose aims did not intersect and/or were at odds); and scholars and PhD students from fields such as criminology, law, public health, cultural, media, feminist, and gender studies, philosophy, and sociology.

Participants were placed into six working groups, and each of these six groups addressed a specific question or topic, with the aim of devising concrete ideas for potential ways to tackle the problem of cyberhate. After the working groups deliberated, they then developed presentations to report their findings to all Symposium participants, and the six presentations together with the discussions/Q&A sessions that followed each presentation were recorded. In the intervening months, we reviewed these recordings, and we compiled the various points that were raised into thematic clusters, which in turn fell into three categories: (i) challenges to recognising, reporting, and tackling cyberhate; (ii) potential solutions to cyberhate; and (iii) twelve stakeholder groups who are involved in the prior two categories.

Although our responses to the AHRC’s consultation questions – and the form and content of this entire submission – are informed by a wide range of sources, a substantial portion of the content and shape of our response to Question 2 (in Section 4.2.) is drawn from the first category of thematic clusters. The third category of thematic clusters informed a substantial portion of our responses to Questions 3, 4, and 5 (respectively, in Sections 4.3., 4.4., and 4.5.). And the influence of the second category of thematic clusters is evident in our responses to all five questions that we address, as well as this whole submission.

By reporting on the findings of the Cyberhate Symposium in this submission, we hope to bring to light a number of important new insights — in particular, about why cyberhate is such an intractably difficult problem to address, but yet what strategies might help to address it.

The Cyberhate Symposium: date, location, aims, design, format, and rationale. The Cyberhate Symposium – formally entitled “Gendered violence online” – was an innovative, hands-on Symposium held at the Red Rattler Theatre in Marrickville, in Sydney, New South Wales, on 7 July 2017. It had three key aims: (i) to steer the conversation about gendered cyberhate away from merely identifying problems and critiquing existing structures, and towards modes of intervention; (ii) to avoid potentially unhelpful “knowledge silos” by facilitating transdisciplinary investigation that included active collaboration between scholars and non-scholars; and (iii) to formulate – and disseminate – potential solutions to misogyny online which were novel and innovative, yet also informed and feasible.

The format was experimental, and designed to be challenging for participants (who were briefed about these aspects of the event in advance). Our rationale was that all too often people find themselves in institutional and personal echo chambers, in which we all end up talking mostly with people with whom we already share views as well as paradigmatic ways of looking at and making sense of the world. In other words, we occupy various online and offline versions of what are known as “filter bubbles” [49].

A great deal of planning was devoted to investigating the best way to structure the event so as to best achieve our stated aims. Eventually it was decided:

- To structure the event as a hands-on workshop rather than presentation-based (so as to maximise the potential for participants to interact with each other rather than simply to deliver “set pieces” they might have delivered in similar versions at other events many times before).
- To ensure participants did not have to engage in lengthy preparation beforehand in order to respect their generosity in devoting a full day of their time to the event.

- To place participants into six, themed working groups: (i) police; (ii) platforms and corporations; (iii) technology design; (iv) law and policy; (v) activism; and (vi) theoretical issues, in order to reduce repetitive discussion and to ensure a wide rather than narrow range of issues were discussed (the members of these working groups can be found in the table below).
- To give each working group a single, key cyberhate-related problem and to ask them to discuss possible solutions to this problem (in an attempt to keep the discussion solutions-focussed).
- To place some participants in groups with themes outside their putative field or area of interest in an attempt to encourage the use of expertise in new contexts and potentially novel ways.
- To include one or more gendered cyberhate targets in each group to enable the perspective of victims to be included in all discussions.
- To inform participants of their working group only on the day of the event (to maximise intra-group dialogue that considered the views of others rather than being overly informed by *a priori* ideas and orientations).

The event was deliberately not staged at a university so as to encourage participation from non-academics. This was partly in response to a number of potential attendees expressing reservations about coming because they were concerned they did not possess adequate tertiary qualifications. At all times, the organisers underlined the value of differing areas of expertise. Jargon was discouraged and a slide reading, “I don’t understand. Could you please explain?” was projected onto a screen behind the stage in order to help participants feel comfortable asking for unfamiliar terms, jargon or complex academic references to be explained. We were gratified that many academics also made use of this invitation to ask colleagues from other disciplines to “please explain” specialised language and concepts.

Outcomes of the event have included media articles in the days after the event, a number of articles-in-progress for academic journals, and this submission. Equally importantly for us in terms of outcomes has been the feedback from participants saying they found the event to have been an excellent learning experience, and that they have been feeding this new knowledge into their own work and professional networks.

On the following four pages we list the participants, their affiliations, and the groups into which they were placed.

SPEAKERS, WORKING GROUPS AND SELECTED LIST OF PARTICIPANTS⁵

Speakers	Mehreen Faruqi (then a Greens NSW MP, now a Senator for NSW) Kath Read (blogger, activist, cyberhate target) Emma A. Jane (cyberhate researcher, University of NSW) Nicole A Vincent (philosopher, Macquarie University)
Guest presenter	Nicole Lawder (Liberal ACT MP)
Guest performers	Maeve Marsden and Libby Wood from Lady Sings It Better

Working group name	Working group members
Law and Policy	<p>Rosalie O’Neale Program Manager, eSafety Women Office of the eSafety Commissioner</p> <p>Daniel Joyce Senior Lecturer, Faculty of Law, UNSW Sydney, Affiliated Research Fellow, Erik Castrén Institute of International Law and Human Rights, University of Helsinki</p> <p>Janin Bredehoeft PhD candidate, Department of Political Economy, University of Sydney</p> <p>Rhys Michie College of Arts & Social Sciences, Australian National University</p> <p>Benjamin Gill Youth Director REELise Incorporated PhD candidate, Brain and Mind Centre, University of Sydney</p> <p>Anjalee de Silva Melbourne Law School University of Melbourne</p>
Police	<p>Carlene Mahoney Inspector, Operational Programs, Major Events and Incidents Group, NSW Police Force</p>

⁵ Please note that this list does not contain the names of all participants since some requested to not be named. Also note that the cited affiliations are shown as they were on the date of the Cyberhate Symposium, however these may have changed in the intervening time.

	<p>Laura Nightingale Senior Sergeant and Legal Consultant Domestic and Family Violence Team Operational Programs NSW Police Force</p> <p>Ezel Jupiter Senior Programs Officer, Domestic and Family Violence Team, Operational Programs, NSW Police Force</p> <p>Mariam Veiszadeh Lawyer, writer, advocate President, Islamophobia Register Australia</p> <p>Tara Moss Author, documentary maker and presenter, speaker, human rights advocate, anti- cyberbullying campaigner, PhD candidate UNICEF ambassador</p> <p>Paloma Brierley Newton Co-founder Sexual Violence Won't Be Silenced</p> <p>Emily Dunn Collective Shout, NSW Volunteer Coordinator</p>
<p>Platforms and Corporations</p>	<p>Rod McGuinness Social Media Manager ABC Radio</p> <p>Kath Read Blogger, librarian, feminist activist</p> <p>Annalise Hartwig Moderator and contributor, Destroy the Joint</p> <p>Nicolas Suzor Associate Professor, School of Law, Queensland University of Technology (QUT)</p> <p>Dominique Coorey Senior technology general manager Telecommunications and finance industries</p> <p>Elena Cama Centre for Social Research in Health, UNSW Sydney</p>

	<p>Rosalie Gillett PhD candidate, School of Justice, Faculty of Law, Queensland University of Technology (QUT)</p> <p>Kathie Melocco Author, speaker, journalist Co-founder, The Respect Campaign</p>
<p>Technology Design</p>	<p>Jordan Newnham Senior Communications Advisor Office of the eSafety Commissioner</p> <p>David Hollingworth Digital Editor, Next Media</p> <p>Gabrielle Nikodem CoderDojo, NSW</p> <p>Melanie Andersen School of Public Health, UNSW Sydney</p> <p>Fiona Andreallo Lecturer USYD, UNSW Sydney, UTS</p>
<p>Theoretical Issues</p>	<p>Amy Gray Writer</p> <p>Chris Fleming Associate Professor Philosophy Research Initiative, School of Humanities and Communication Arts, University of Western Sydney</p> <p>Amanda Elliot Department of Sociology and Social Policy, University of Sydney</p> <p>Jocelyn Hungerford Freelance editor and writer</p> <p>Son Vivienne Lecturer in Digital Media, Flinders University of South Australia</p> <p>Kerryn Drysdale Department of Gender and Cultural Studies, University of Sydney Centre for Social Research in Health, UNSW Sydney</p>

	<p>Lucy Hackworth Gender studies scholar (online harassment) Gender Erasmus Mundus Masters (GEMMA) Program Utrecht, The Netherlands</p> <p>Jackie McMillan Policy, Media & Communications Officer, Sex Workers Outreach Project</p>
<p>Activism</p>	<p>Jessamy Gleeson 'Cherchez La Femme' producer, SlutWalk Melbourne organiser, PhD candidate, Department of Media & Communication, Swinburne University of Technology</p> <p>Anthony Minniecon Programs Manager KWY Aboriginal & Torres Strait Islander Family Services, Adelaide</p> <p>Elinor Lloyd-Philipps Lingerie blogger @The Nylon Swish</p> <p>Caitlin McGrane Writer, marketing communications professional, social media researcher</p> <p>Carly Pettiona School of Languages and Linguistics University of Melbourne</p> <p>Lucy Le Masurier Sexual Violence Won't Be Silenced</p>
<p>Other participants</p>	<p>Maliha Aqueel Policy advisor, Parliament of NSW PhD Student, University of Sydney</p> <p>Teresa Avila Cofounder, Chair, Treasurer – Red Rattler Theatre</p> <p>Ariadna Matamoros-Fernández PhD candidate, Digital Media Research Centre (DMRC), Queensland University of Technology (QUT)</p>

2.2.3. *Prior work*

The third source of arguments and data that has informed the shape and content of this submission is the ongoing series of research projects in which Dr Jane, frequently in collaboration with Dr Vincent, has mapped and studied the history, manifestations, nature, prevalence, aetiology, and consequences of gendered cyberhate. Dr Jane has also used approaches from internet historiography to archive – starting in 1998 and then on an ongoing basis – many thousands of reports or incidences of gendered cyberhate in many domains.⁶ We list some of the outputs of these research projects below, and note that some of them may also appear in the References section.

Jane, Emma A. (forthcoming 2019), “Hating 3.0 and the Question of Whether Anti-Fan Studies Should Be Renewed for Another Season”, in Click, Melissa (ed.), *Dislike, Hate, and Anti-Fandom in the Digital Age*. New York: New York University Press.

Vincent, Nicole. A & Jane, Emma A. (2018), “Cognitive Enhancement: A Social Experiment with Technology”, in van de Poel, Ibo, and Asveld, Lotte, and Mehos, Donna C. (eds.), *New Perspectives on Technology in Society: Experimentation Beyond the Laboratory*. Oxon & New York: Routledge, pp. 125-14.

Jane, Emma A. (2018), “Gendered Cyberhate as Workplace Harassment and Economic Vandalism”, *Feminist Media Studies*, special edition on Online Misogyny. DOI: 10.1080/14680777.2018.1447344.

Jane, Emma A. (2017), *Misogyny Online: A Short (and Brutish) History*. LA, London & New Delhi: SAGE.

Jane, Emma A. (2017), “Systemic Misogyny Exposed: Translating Rapeglish From the Manosphere With a Random Rape Threat Generator”, *International Journal of Cultural Studies*. DOI: 10.1177/1367877917734042.

Jane, Emma A. (2017), “Feminist Digilante Responses to a Slut-Shaming on Facebook”, *Social Media + Society*, April-June. DOI: 10.1177/2056305117705996.

Jane, Emma A. (2017), “Online Misogyny and Feminist Digilantism”, *Continuum: Journal of Media & Cultural Studies*. DOI: 10.1080/10304312.2016.1166560.

⁶ For information about Dr Jane’s ongoing research, her collaborations with Dr Vincent, and Dr Vincent’s other work that contributed to this submission, please see Section 3.2.3. below.

- Jane, Emma A. & Vincent, Nicole A (2017), Random Rape Threat Generator (original creative work). 18 January. Accessed from <https://www.rapeglish.com/> and <http://www.rapethreatgenerator.com/> on November 10 2018.
- Jane, Emma A. (2017), "Gendered Cyberhate: A New Digital Divide?", in Ragnedda, Massimo and Muschert, Glenn W. (eds.). *Theorizing Digital Divides*. Oxon: Routledge, pp. 158-198.
- Jane, Emma A. (2017), "Feminist Flight and Fight Responses to Gendered Cyberhate", in Marie Segrave, Marie and Vitis, Laura (eds.), *Gender, Technology and Violence*. Oxon: Routledge, pp. 45-61.
- Jane, Emma A. (2017), "Gendered Cyberhate, Victim-Blaming, and Why the Internet Is More Like Driving a Car On a Road Than Being Naked in the Snow", in Martellozzo, Elena and Jane, Emma J. (eds.), *Cybercrime and its Victims*. Oxon: Routledge, pp. 61-78.
- Jane, Emma A. and Martellozzo, Elena (2017), "Introduction: Victims of Cybercrime on the Small "i" Internet", in Martellozzo, Elena and Jane, Emma J. (eds.), *Cybercrime and its Victims*. Oxon: Routledge, pp. 1-24.
- Vincent, Nicole A and Jane, Emma A. (2017), "Beyond Law: Protecting Cyber Victims Through Engineering and Design", in Martellozzo, Elena and Jane, Emma J. (eds.), *Cybercrime and its Victims*. Oxon: Routledge, pp. 209-223.
- Vincent, Nicole A (2017), "Victims of cybercrime: Definitions and challenges", in Martellozzo, Elena and Jane, Emma J. (eds.), *Cybercrime and its Victims*. Oxon: Routledge, pp. 27-42.
- Jane, Emma A. and Vincent, Nicole A (2017), "Why We Need to Get Comfortable with Reporting Cyberhate", UNSW Sydney Arts & Social Sciences Newsroom, 19 July. Accessed from <https://www.arts.unsw.edu.au/newsroom/articles/why-we-need-to-get-comfortable-with-reporting-cyberhate/> on 10 November 2018.
- Vincent, Nicole A and Jane, Emma A (2017), "A Crime Is a Crime, Even If It's Online - Here Are Six Ways to Stop Cyberhate", ABC News, 18 July. Accessed from <http://www.abc.net.au/news/2017-07-18/six-ways-to-stop-cyberhate/8721184> on 10 November 2018.
- Jane, Emma A. and Vincent, Nicole A (2017), "Women Online Are Getting Used to Cyber Hate. They Need to Get Used to Reporting It", *The Sydney Morning Herald*, 18 July. Accessed from <http://www.smh.com.au/lifestyle/news-and-views/opinion/women-online-are-getting-used-to-cyber-hate-they-need-to-get-used-to-reporting-it-20170717-gxctr8.html> on 10 November 2018.

- Jane, Emma. A (2017), "What the Random Rape Threat Generator Tells Us About Online Misogyny", Women's Media Center Speech Project, 18 January. Accessed from <http://wmcspeechproject.com/2017/01/18/what-the-random-rape-threat-generator-tells-us-about-online-misogyny/> on 10 November 2018.
- Jane, Emma A. (2017), "Rapeglish – the Hyperbolic, Sexualised Vitriol Found Online", The Sydney Morning Herald, 7 January. Accessed from <http://www.smh.com.au/lifestyle/news-and-views/opinion/rapeglish--the-hyperbolic-sexualised-vitriol-found-online-20170106-gtmwgl.html> on 10 November 2018.
- Jane, Emma A. (2016), "Rapeglish: A Program that Spits Out Hate – For the Greater Good", Social Science Space, 28 December. Accessed from <http://www.socialsciencespace.com/2016/12/rapeglish-program-spits-hate-greater-good/> on 10 November 2018.
- Jane, Emma A. (2016), "Stopping Online Abuse Isn't Censorship: It's the Least We Can Do", The Age, 14 July. Accessed from <http://www.theage.com.au/comment/stopping-online-abuse-isnt-censorship-its-the-least-we-can-do-20160713-gq4oj5> on 10 November 2018.
- Jane, Emma A. (2016), "DIY Internet Justice is a Symptom, Not a Solution to Online Misogyny", Daily Life, 11 April. Accessed from <http://www.dailylife.com.au/news-and-views/dl-culture/diy-internet-justice-is-a-symptom-not-a-solution-to-online-misogyny-20160410-go2z6z.html> on 10 November 2018.
- Jane, Emma A. (2016), "What Bit About the Wrongs of Sexual Threats Against Women Do Courts and Men Not Get?", The Conversation, 4 August. Accessed from <https://theconversation.com/what-bit-about-the-wrongs-of-sexual-threats-against-women-do-courts-and-men-not-get-63447> on 10 November 2018.
- Jane, Emma A. (2015), "Rape Threats and Cyberhate? Vote No to the New Digital Divide", The Conversation, 22 June. Accessed from <https://theconversation.com/rape-threats-and-cyberhate-vote-no-to-the-new-digital-divide-43388> on 10 November 2018.
- Jane, Emma A. (2015), "How to Keep the Internet Hot", Medium, 28 August. Accessed from <https://medium.com/festival-of-dangerous-ideas/how-to-keep-the-internet-hot-6fb9a37ad120#.mbrp5fybk> on 10 November 2018.
- Jane, Emma A. (2015), "What I've Learned From My Study Into Gendered Cyberhate", Daily Life, 31 August. Accessed from <http://www.dailylife.com.au/news-and-views/dl-culture/what-ive-learned-from-my-study-into-gendered-cyberhate-20150828-gj9qsu.html> on 10 November 2018.

Jane, Emma A. (2015), "Flaming? What Flaming?: The Pitfalls and Potentials of Researching Online Hostility", *Ethics and Information Technology*, 17(1): 65-87. DOI: 10.1007/s10676-015-9362-0.

Vincent, Nicole A and Jane, Emma A. (2014), "Put Down the Smart Drugs – Cognitive Enhancement is Ethically Risk Business", in *The Conversation* (ed.). 2014: A Year in the Life of Australia. Sydney: Future Leaders, pp. 120-124.

Jane, Emma A. (2014), "'Back to the Kitchen, Cunt': Speaking the Unspeakable About Online Misogyny", *Continuum – Journal of Media & Cultural Studies*, 28(4): 558-570. DOI: 10.1080/10304312.2014.924479.

Jane, Emma A. (2014), "Beyond Antifandom: Cheerleading, Textual Hate and New Media Ethics", *International Journal of Cultural Studies*, 17(2): 175-190. DOI: 10.1177/1367877913514330.

Jane, Emma A. (2012), "'Your a Ugly, Whorish, Slut' – Understanding E-Bile", *Feminist Media Studies*, 14(4): 531–546. DOI:10.1080/14680777.2012.741073.

2.3. AUTHORS

Dr Emma A. Jane is regarded as one of the world's most authoritative researchers on gendered digital harassment and its impacts. Her transdisciplinary work on gendered cyberhate and digital citizenship has been acknowledged as generating "a great deal of impact internationally and in policy and regulatory settings" [70]. Blockchain, digital disruption, digital literacy, the future of work, cyberhate, digilantism, and value-sensitive design, are some of the foci of her ongoing research into the social and ethical implications of emerging technologies.

During the first 25 years of her professional life, Dr Jane achieved national prominence as a senior journalist, columnist, and feature writer at the Sydney Morning Herald and the Australian newspapers. She publishes prolifically, and has written nine books, 24 chapters, and 15 journal articles, as well as co-editing a recent collection on cybercrime. She has given 75 talks at events in Australia and abroad, and received a total of 20 awards, prizes and grants in recognition of the excellence of her scholarly research, her journalism, and her fiction-writing. In 2016, the public benefit of her research into misogyny online was recognised when she was named the Anne Dunn Scholar of the Year. This followed her receipt, in 2014, of the aforementioned Discovery Early Career Researcher Award (DECRA) from the Australian Research Council (ARC) to fund a three-year research project into gendered cyberhate. Most recently, in 2017, she received the Dean's Award for Achievements by an Early Career Researcher awarded by her Faculty at UNSW, Sydney.

Dr Jane has presented the findings of her research to the Australian Human Rights Commission and the Australian Government's Workplace Gender Equality Agency, and regularly speaks at large, public events such as the Festival of Dangerous Ideas and the All About Women festivals at the Sydney Opera House. Since 2016, she has also been an editorial board member of the *International Journal of Cultural Studies*. Previously she served for five years on Australia's Advertising Standards Board. She is frequently interviewed about her research by international media outlets, and has given 80 interviews since 2014 alone.

Dr Jane has a demonstrated track record of forging strategic collaborations with industry, government, and NGOs. In late 2016 and early 2017, for instance, she collaborated with Soraya Chemaly from the US-based Women's Media Center (WMC) Speech Project and Camille Francois from Google Jigsaw as part of these organisations' efforts to address the problem of gendered cyberhate. She provided input on how her research data bases of misogynist cyberhate might be used to help develop machine-learning tools better able to detect and respond to gendered threats and abuse. In 2017, Dr Jane – along with her principle research collaborator Dr Vincent – organised the aforementioned Symposium aimed at generating solutions for gendered cyberhate. The conference's 60 participants included the NSW Legislative Council MP Mehreen Faruqi, the Liberal member for Brindabella Nicole Lawder, and the ABC's Social Media "Self-Defence" Manager Rod McGuinness, as well as representatives from the Office of the eSafety Commissioner, the NSW Police Force, and domestic and family violence organisations. In the aftermath of the Symposium, the NSW parliament unanimously passed a motion congratulating Dr Jane and Dr Vincent on the event.

Dr Jane's latest monograph, *Misogyny Online: A Short (and Brutish) History*, has received excellent reviews. The international journal *Information, Communication & Society* applauds its "winning combination of conceptually and philosophically rich analysis, forensic and details-oriented storytelling" and describes it as "essential reading for those working in the field" [70]. The UNICEF National Ambassador for Child Survival Tara Moss, meanwhile, praises the book as "a rigorous, necessary and at times terrifying exploration of one of the most pressing and rapidly growing forms of harassment and abuse of women and girls today."

Dr Nicole A Vincent has taught, written, and delivered talks on a wide range to topics including gendered cyberhate and cybercrime, transgender-related public policy, law and neuroscience, smart drugs (aka cognitive enhancement), biomedical moral enhancement, free

will and determinism, and how emerging technologies – such as gene editing, blockchain, and autonomous vehicles – can foster human flourishing.

The concept of responsibility occupies centre stage in Dr Vincent's scholarly pursuits, and socially responsible innovation and value-sensitive design are key features of her approach to a wide range of topics in a diverse range of fields including neuroethics, neurolaw, bioethics, philosophy and ethics of emerging technologies, political philosophy, public policy, jurisprudence and philosophy of law, bioethics, media, feminism, gender, and happiness.

Dr Vincent's research has been funded by more than \$1 million external grants. She has published 40 peer reviewed articles, delivered almost 100 academic talks, and organised 19 conferences. She also talks regularly about her work on television and radio – as well as in a wide range of other public contexts – to ensure that her research is influenced by and useful to society. For example, in 2014 she delivered a TEDxSydney talk at the Sydney Opera House on the ethics of “smart drugs”, and participated in an Intelligence Squared debate at Angel Place arguing against the proposition that society would necessarily flourish under female rule.

Since obtaining her PhD in philosophy from the University of Adelaide in 2007 and before joining UTS, she also held positions at the University of Auckland in New Zealand, Technische Universiteit Delft in The Netherlands, Georgia State University in USA, Macquarie University, Charles Sturt University, and UNSW Sydney.

2.4. DISCLAIMERS

Dr Emma A. Jane is employed by UNSW Sydney and, from 2015 to 2017, was the recipient of an Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA Project ID: DE150100670).

Dr Nicole A Vincent is employed by University of Technology Sydney (UTS), and she recently co-led the development of UTS's institutional response to the AHRC's HRT Issues Paper. Dr Vincent is also an Honorary Fellow and Affiliate Member of the Centre for Agency, Values, and Ethics in the Department of Philosophy at Macquarie University.

However, the views expressed in this submission are those of the authors, and not necessarily those of the ARC, UNSW Sydney, UTS, or Macquarie University.

2.5. LIMITATIONS

Because of the focused nature of this submission, which concentrates on the topic of cyberhate, which in turn involves technologies that underpin a range of internet services, the chief limitation of this submission is that it only addresses the first five questions.

Questions 6-8 have a specific focus on Artificial Intelligence Informed Decision Making (AIDM), and although in response to Question 5 we touch on this topic, our assessment is that at present AIDM's ability to protect human rights from cyberhate is limited, which explains why we have little further to add on this topic in Questions 6-8.

Questions 9 and 10 have a specific focus on disability, and although people with disabilities are also among the vulnerable groups that are disproportionately the targets of cyberhate, our discussion has nothing specific to say about this particular vulnerable population. Since our submission has taken a fairly broad approach in this regard to defining the scope of which populations fit under the umbrella heading of "vulnerable groups", offering answers to Questions 9 and 10 would have detracted from our core message which is that many vulnerable groups are targeted by cyberhate. Naturally, different groups are picked on for different reasons, and they experience different degrees of exposure to – and suffer different harms from cyberhate. However, since the topic of Dr Jane's DECRA research fellowship focused on *gendered* cyberhate, apart from our discussion in Section 4 which has a distinct focus on the way that cyberhate impacts on women, we felt that it would not be appropriate to venture specific comments about other vulnerable groups.

3. CYBERHATE

Cyberhate is a critical human rights issue that is inextricably linked to a range of internet technologies, which disproportionately affects women, girls, LGBTIQ+ people, and a range of racial, cultural, and linguistic minorities, and it requires urgent attention. This key claim underpins our responses in Section 4 of this submission to the consultation questions listed in the AHRC's HRT Issues Paper.

We are heartened by the fact that some of the harms, vulnerable populations, and technologies discussed in the present submission are also noted in the AHRC's HRT Issues Paper. For instance, Section 4.2 of the latter states:

[N]ew technology and online platforms present enormous opportunities to advance gender equality and are a powerful tool for women to increase their access to education and information, social connectedness and improve their economic security. However, women are also disproportionately the target of personal, sexual and gender-based cyber abuse.

A 2016 study found that 76% of women under 30 years of age, have reported experiencing online harassment, and almost half (47%) of all women had been targets. Similarly, one in four lesbian, bisexual and transgender women report targeted sexual orientation harassment. More recent research on the experiences of women in Australia found that, of those that had experienced online abuse and harassment, 42% of women said it was misogynistic or sexist in nature, and 20% said it had included threats of physical or sexual violence.

The social and economic consequences of widespread automation are also likely to be different for women than men, with significant implications for socio-economic equality and the global gender gap. The disparity in global access to technology and the internet may also have detrimental consequences for women, particularly for future economic opportunities. [58: 20]

We applaud the AHRC's acknowledgement of these issues, and we further note that this aligns with the recently articulated positions of the United Nations (UN) [13] and Amnesty International [14] [57], who also recognise that cyberhate violates human rights.

However, despite the growing recognition of cyberhate as a critically important human rights issue – along with a growing body of international research showing both its prevalence and the harms that it causes⁷ – at the same time there also exists an abundance of media and other commentary claiming that cyberhate is *not* a serious problem because it involves words

⁷ See, for instance [5], [13], [14], [16], [17], and [21].

rather than actions, and only virtual domains rather than “real” life.⁸ If these sentiments were to be stated in colloquial terms, they would amount to something in the order of the claim that cyberhate is not of serious concern since it is, after all, “just words” and/or “just the internet”.

This gross and surprisingly difficult-to-dislodge misperception and mischaracterisation of cyberhate has come about for two closely-related reasons. *Firstly*, a number of insidious structural, institutional, and technological factors that we discuss in Section 4.2 of this submission continue to make it difficult for cyberhate to even be recognised in certain important contexts. *Secondly*, and as a direct consequence of the first reason, there persists a despairingly common under-appreciation of precisely what cyberhate involves, and of its deeply damaging effects. Together, these two reasons explain why, despite the AHRC’s, the UN’s, and Amnesty International’s recognition that cyberhate raises human rights concerns, in some circles this view remains controversial, and is even outright rejected.

In light of the above, the purpose of this section is to set the record straight by pre-emptively responding to those who might doubt the critical importance of cyberhate as a human rights issue, owing to their lack of knowledge of what these offences involve, and of what factors account for the continuing lack of data about cyberhate. By doing this we also hope that those to whom the AHRC will make its own further recommendations will also recognise cyberhate as a genuine threat to human rights, and that they too will acknowledge that this threat stems in important ways from a constantly evolving new technology (the internet) that unfortunately keeps evolving in ways that, if left un-addressed, will continue to leave people vulnerable to exposure to cyberhate. If this recognition does not occur, our concern is that the threats to human rights that we discuss (and the opportunities to address them) in Section 4 will not be recognised nor deemed as sufficiently important to warrant being classified as genuine threats to human rights. For this reason, we aim to offer the most robust support possible for the claim that the cyber abuses which we call “cyberhate” should be included within the boundaries of the AHRC’s existing concerns about human rights and technology.

In this section we offer a brief history of cyberhate, as well as outline – via a combination of empirical data, case studies, and uncensored examples – cyberhate’s signal characteristics. Our aim is to paint an uncomfortably unambiguous image of what cyberhate is, and to recite the wide range of tangible harms to important interests of real victims for which it is responsible. The reality and significance of these victims, harms, and interests is what grounds

⁸ For a lengthy discussion of this topic, please see [5: 76-87].

our claim that cyberhate is a critical human rights issue, and the fact that these things spring from a new and constantly updated technology – a technology that never seems to get the right kind of update to protect the vulnerable – underpins our claim that this is a human rights issue of high relevance to the topic of AHRC’s HRT Issues Paper.

3.1. The rise of cyberhate

Gendered cyberhate was relatively rare and mild in the early decades of the internet. It has, however, become far more prevalent, visible, noxious, and directly threatening since at least 2010 [5: 16-42]. These amplifications are likely a flow-on effect from the self-publishing and networking opportunities associated with what is known as the Web 2.0 era. “Web 1.0” is generally used to describe those early decades of the internet when content was mostly static and delivered in a read-only format. “Web 2.0” refers to the shift – most obvious from around 2006 – towards user-generated material, interactivity, collaboration and sharing. Put simply, the Web 2.0 era has given online antagonists access to targets (and, unfortunately, appreciative audiences) in a way that was not previously possible.

The abuse and harassment of women online typically involves sexually explicit invective, hyperbolic yet plausible rape and death threats, and/or persistent, unwanted sexual advances from senders who often become hostile if ignored or rebuffed [5: 16–18]. The discourse involved frequently: passes scathing and explicit judgement on women’s appearance, sexual attractiveness and/or perceived sexual activeness; deploys *ad hominem* invective; is couched in terms involving hyperbolic misogyny, homophobia, and/or sexually graphic imagery; and prescribes coerced sex acts as all-purpose correctives. Increasing numbers of women are also reporting instances of: cyberstalking; rape blackmail videos; large groups attacking individuals (sometimes with the explicit purpose of causing job loss or career derailment); malicious impersonation; “sextortion” (the blackmailing of targets in order to extort them to perform sexual acts online); revenge porn (the non-consensual uploading of sexually explicit material of a subject without their consent); and “doxing” (the publishing of personally identifying information, usually to incite internet antagonists to hunt targets in “real” life) [5: 34–5].

Abuse and harassment is frequently image- as well as text-based. Photo manipulation, for example, is often used to place an image of a target into a scene involving sex and/or violence. An attack on the tech designer Kathy Sierra included doctored photos depicting her being choked by undergarments, and with nooses next to her head [26]. The feminist cultural critic Anita Sarkeesian, meanwhile, has received countless images of men ejaculating onto her photo

[27]. One man went so far as to create an online game called “Beat Up Anita Sarkeesian” in which players could “punch this bitch in the face” until Sarkeesian’s face became bloody and battered [28]. It has also become common practice for men to send unsolicited and unwanted photos of their genitals – aka “dick pics”. According to 2018 research by YouGov UK, four in 10 female millennials have been sent an unsolicited penis photo [29].

The YouGov UK study is just one of many which demonstrate that cyberhate is extremely prevalent as well as noxious. In November, 2017, Amnesty International UK published research showing that one in five women in the UK, the US, New Zealand, Spain, Italy, Poland, Sweden, and Denmark had experienced online abuse or harassment. Of these: more than a quarter (27 per cent) received direct or indirect threats of physical or sexual violence; almost half (47 per cent) had experienced sexist or misogynistic abuse; and one-third (36 per cent) felt their physical safety had been threatened [14]. This followed an earlier report by the UN Broadband Commission stating that 73 per cent of women and girls had been exposed to or had experienced some form of online violence [13]. The report acknowledged: that women were 27 times more likely to be abused online than men; that 61 per cent of online harassers were male; and that women aged between 18 and 24 were at particular risk [13: 15]. With growth in global connectivity and the internet’s ever-increasing penetration into daily life, the UN warned that, unchecked, cyber violence against women and girls (“cyber VAWG”) risked becoming “a 21st century global pandemic with significant negative consequences for all societies in general and irreparable damage for girls and women in particular” [13: 2]. These international figures comport with data collected in individual countries. For instance, a survey of 3,000 Australians aged 18 to 54 revealed that one in five women overall and two in five women aged 18 to 19 report having been targeted for digital sexual harassment [17: 1-2].

3.2. Gendered characteristics

As we elaborate in great detail in Section 4.2. of this submission, uncensored examples of cyberhate are critically important for demonstrating its nature and revealing its devastating impact. We thus begin this section with two recent Australian examples:

Annie Nolan: In 2015, the Melbourne-based blogger Annie Nolan was waiting for a train with her two-year-old twins when she wrote a list of what were intended to be funny statements on the back of two large envelopes. These statements took the form of a series of answers to questions strangers frequently asked her about her twins. They included: “YES, THEY ARE MINE”, “NO, NOT IDENTICAL”, “YES, I KNOW THEY LOOK ALIKE THOUGH”, and “YES, MY HANDS ARE FULL (SOMETIMES WITH 2 GLASSES OF WINE JUST TO GET THROUGH”

[15]. Nolan uploaded the photo to her blog's public Facebook page when the latter had only about 200 followers. The post, however, went viral and was viewed more than two million times over two days. Antagonists inundated Nolan with abuse and death threats, including someone saying they would shove a broken glass into her face if they saw her on the street [7]. Others bombarded her with photos of their dead children and the urns of their cremated babies alongside messages accusing her of being a bad mother ungrateful that her children were alive.

Waleed Aly: Waleed Aly – one of the hosts of the popular Australian current affairs program *The Project* – has said he deliberately avoids social media because he does not want to worry about “what Twitter is going to say” when he articulates his political positions on air [8]. Antagonists are, however, still able to target Aly for abusive tweets via the hashtag #waleedaly, and via his family and colleagues. One man, for instance, directed a #waleedaly tweet at Aly's spouse, the academic Susan Carland, as well as to the Australian Muslim lawyer Mariam Veiszadeh (both of whom *do* hold Twitter accounts). The tweet read: “This is my death threat to you and Waleed and his hijabi scumfuk floozie. I hope you all meet with natural accidents” [9]. Others have used online platforms to call Aly a “RACIST Muslim CUNT” [10], and a “muslime cunt”, and “Muzzie prick” [11].

These examples may convey the impression that cyberhate is equally a problem for women and men. However, although women and men in Australia are indeed equally likely to report experiencing digital harassment and abuse, women are more likely to report sexual harassment, are significantly more likely to be “very or extremely upset” by the abuse, and are more likely to take actions such as changing their online details or profile settings, or leaving a site [17: 1]. Similarly, a 2014 study by the Pew Research Center in the United States reports that men are more likely to experience name-calling and embarrassment – harassment of the types categorised as less severe: “a layer of annoyance so common that those who see or experience it say they often ignore it” [21: 2-3]. Young women, in contrast, are particularly vulnerable to severe types of abuse such as stalking, and sexual harassment. University of Maryland researchers have found that internet accounts with feminine usernames incur an average of 100 sexually explicit or threatening messages for every four received by male users [16: 14]. The Waleed Aly example above also shows the way that cyberhate aimed at men can still involve misogyny in that male targets are frequently attacked via their female family members, friends, and colleagues [23: 565].

Gendered cyberhate can be contextualised within a broader “pandemic” of gendered violence – as per data showing that one in three women experience physical or sexual violence over the course of their lifetimes [13: 2]. It manifests in a wide variety of practices which can

be situated along a spectrum of violence and harm depending on the context. A real-life example which sits at the most extreme end of the spectrum is that of Jebidiah James Stipe, a 28-year-old American man who impersonated his former female partner on the internet site Craigslist and published a photo of her alongside text saying she was seeking “a real aggressive man with no concern for women” [16: 5]. More than 160 people responded to the ad, including a man who – after Stipe divulged his ex-partner’s address – arrived at the woman’s home, bound and blindfolded her, and raped her at knifepoint [16: 5-6]. Both Stipe and the rapist were subsequently jailed for 60 years to life in prison [22].

Cyberhate also includes direct and credible threats of violence. For example, when the British Labour MP Stella Creasy spoke in support of a student feminist activist who had campaigned to have Jane Austen replace Charles Darwin on the English £10 note, she received a tweet reading, “YOU BETTER WATCH YOUR BACK... IM GONNA RAPE YOUR ASS AT 8PM AND PUT THE VIDEO ALL OVER THE INTERNET” [23]. Consider, too, the response to journalist Sady Doyle’s suggestion that gendered harassment had become an inevitable consequence of blogging while female [24]. “**Simply put,**” the men’s rights activist Paul Elam wrote on his website A Voice for Men, “**we are coming for you. All of you.** And by the time we are done you will wax nostalgic over the days when all you had to deal with was someone expressing a desire to fuck you up your shopworn ass” [25, emphasis in original].

During the vicious mob attacks on women in 2014 dubbed “GamerGate”, the games developer Zoë Quinn accumulated 16 gigabytes of abuse [43]. Her anonymous antagonists circulated her home address and personal photos online [44] and her Wikipedia entry was edited to read: “Died: soon.” After this entry was deleted, a new one appeared reading: “Died: October 13, 2014” – the date of her next scheduled public appearance [45]. Harassers threatened Quinn’s father, and the future employers of her new boyfriend, who subsequently had a pending job offer withdrawn [43]. Quinn was inundated with threats such as, “Im not only a pedophile, ive raped countless teens, this zoe bitch is my next victim, im coming slut”, and “kill yourself. We don’t need cunts like you in this world” [43]. (Incitements to suicide are a common cyberhate tactic, particularly when subjects are known to suffer from mental illness. It was public knowledge, for instance, that Quinn experienced chronic depression [43]. For a further discussion of this issue, see the example of the late Australian celebrity Charlotte Dawson discussed in Section 2.2.3. below.)

The noxiousness and apparent credibility of such threats has led some cyberhate targets to leave their homes in fear. Quinn, for instance, fled her home shortly after the assault on her

began [45]. After having received such a massive quantity of abuse, her concern was that it would only be a matter of time before one of her anonymous critics eventually made good on their threats to kill her [45]. In the same month – August, 2014 – Sarkeesian also left her home after receiving a series of graphic death threats which demonstrated knowledge of her and her parents’ home addresses [46]. Shortly after this, Sarkeesian cancelled a speaking event at Utah State University after an anonymous emailer threatened “the deadliest school shooting in American history” if her talk went ahead as planned [46]. This email said Sarkeesian was “everything wrong with the feminist woman” and would “die screaming like the craven little whore that she is” [46]. Around the same time, the personal details of the American games designer Brianna Wu – the co-founder of the Boston game studio Giant Spacekat – were posted on the 8chan web site, and within minutes someone tweeted at her saying, “I’ve got a K-bar⁹ and I’m coming to your house so I can shove it up your ugly feminist cunt” [44]. Wu also left her home because she feared for her safety. Her observation was that this was “not just casual sexism”, it was “angry, violent sexism ... Every woman I know in the industry is scared. Many have thought about quitting” [44].

3.3. Impacts and harms

Emerging research is shedding light on the profound, multi-faceted suffering that can be experienced by cyberhate targets.¹⁰ Dr Jane’s studies, for instance, are showing the way the coercive force of gendered cyberhate is causing women significant emotional, social, financial, professional, and political harm. It is constraining their ability to find jobs, market themselves, network, engage politically, socialise, and partake freely in the sorts of self-expression, self-representation, creativity, interactivity, and collaborative enterprises celebrated as key benefits of the Web 2.0 era. Harassment and threats at the most extreme end of the spectrum can cause women to experience debilitating fear, trauma, and life disruption. Indeed, some women have developed mental health problems or experienced breakdowns.

A particularly infamous example of cyberhate involves the British activist Caroline Criado-Perez whose 2013 campaign to have the Bank of England review its decision to have an all-male line-up on bank notes resulted in her receiving tweets such as “KISS YOUR PUSSY GOODBYE AS WE BREAK IT IRREPARABLY”, and “If your friends survived rape they

⁹ Our reading of “K-bar” here is that it is a misspelling of “ka-bar” – a combat knife.

¹⁰ See, for example: [16], [30], and [31].

weren't raped properly" [5: 1-2]. During the height of the attack against her – a time in which she was receiving around 50 abusive and threatening messages per hour – Criado-Perez says:

The immediate impact was that I couldn't eat or sleep. I lost half a stone in two days. I was just on an emotional edge all the time. I cried a lot. I screamed a lot. I don't know if I had a kind of breakdown. I was unable to function, unable to have normal interactions.

An example from Australia in 2012 involves the depressed TV presenter Charlotte Dawson who was hospitalised after receiving a barrage of messages such as: "Freedom of speech, you fucking bimbo? Go kill yourself"; "I speak for everyone in the universe. Bitch, you need to kill yourself"; and "Go kill yourself you fucking whore" [48]. Dawson was still engaging with her online attackers – tweeting "Hope this ends the misery" and "You win" – only an hour before an ambulance was called to her home in Sydney because of a suicide attempt [47] [48]. Eighteen months later, the former *Australia's Next Top Model* judge took her own life. While a consideration of Dawson's mental health in general terms is obviously relevant, it is difficult to avoid the conclusion that her online experiences were intimately bound up with her self-destructive behaviour as proximate if not ultimate causes.

Returning to the example of Annie Nolan outlined in Section 2.1.1, Nolan says that she believes she might have taken her own life if she had received this sort of abuse when she was 19 or 20:

Anyone that is ... even a tiny bit emotionally fragile or not okay with themselves... [might] ... hurt themselves. Because the things that are said about you are just utterly awful ... it really is such an extreme feeling. [7]

Nolan says the abuse also caused her friends and family a great deal of distress because they were worried that – given Nolan's past history of being bullied severely at high school – it might be too much for her to endure. Nolan's mother, in particular, was "really traumatised" and lost two or three kilos of weight during the incident [7]. Accounts such as these comport with the case made by the Australian academics Nicola Henry and Anastasia Powell that harms in the supposedly "virtual" world can have real bodily and psychological effects, and "at least as much impact on a person as traditional harms occurring against the physical body" [33].

3.4. Workplace harassment

Much of the cyberhate women receive at work involves abuse and harassment which would be in clear breach of various workplace-related regulations and guidelines if it occurred in offline contexts. This material often involves content and arrives in forms that constitute a form of

workplace harassment and/or economic vandalism. The latter is a term we use to encapsulate a range of professional and economic harms which result from the receipt of gendered cyberhate and which do not occur in contexts that can neatly be captured by the term “workplace harassment”.¹¹ A specific example would be a woman who is fired, demoted, or passed over for a job interview because an employer or potential employer searches her name online and discovers intimate photographs and footage of her that have been posted by a vindictive former partner.

This dimension of the cyberhate problem is particularly acute for women who work in the media. In 2016, The Guardian engaged in a quantitative analysis of its own comment threads and – after examining 70 million remarks – found that of the 10 regular writers who received the most abuse, eight were women (four white and four non-white) while two were men of colour [18]. The 10 regular writers who received the least abuse were all men. Since 2010, journalism produced by female contributors has consistently attracted more comments requiring blocking by The Guardian’s moderators, with articles about feminism and rape attracting very high levels of blocked comments (as opposed to comments on articles about crosswords, cricket, horse racing, and jazz, which tend to be ‘respectful’) [18]. The largest number of objectionable comments targeted Jessica Valenti, the feminist writer and founder of the blog Feministing [50]. Similar trends have been observed in Australia, where a 2016 survey by the Women in Media group found that 41 per cent of female journalists were being harassed, bullied, or trolled online [51].

Three months after The Guardian study was published, Valenti withdrew from social media after the attacks against her online were extended to her child. Tweeting about the decision, she wrote: “This morning I woke up to a rape and death threat directed at my 5 year old daughter. That this is part of my work life is unacceptable” [52]. The following year, the American writer and performer Lindy West also quit Twitter [53], having previously said she constantly felt the pull to change careers because of the exhaustion of spending years dealing with workplace harassment equivalent to “hundreds of men popping into your cubicle in the accounting department of your mid-sized, regional dry-goods distributor to inform you that—hmm – you’re too fat to rape, but perhaps they’ll saw you up with an electric knife?” [54]. As she put it:

¹¹ See, for instance, see [5: 67-8, 97-8, 116] and [34].

People who don't spend much time on the internet are invariably shocked to discover the barbarism—the eager abandonment of the social contract—that so many of us face simply for doing our jobs. [54]

Online abuse has the power to destroy women's reputations in ways which have significant and ongoing repercussions for their future employment prospects. Findings from the Pew Research Center show that of those people targeted for physical threats and sustained harassment, about a third feel their reputations have been damaged [21: 7]. The legal scholar Danielle Keats Citron's work shows that schools have fired teachers whose naked photos have appeared on revenge porn sites, while a government agency terminated a woman's employment after a co-worker circulated her nude photograph to colleagues [55]. Cogent, too, is the fact that most employers rely on candidates' online reputation to filter applicants. Nearly 80 per cent of employers consult search engines to collect intelligence on job applicants, and about 70 per cent of applicants are rejected because of these findings [56]. Common reasons for not interviewing and hiring applicants include concerns about “lifestyle”, “inappropriate” online comments, and “unsuitable” photographs, videos, and information. As Citron puts it:

The simple but regrettable truth is that after consulting search results, employers don't call revenge porn victims to schedule an interview or to extend offers. It's just seen as good business to avoid hiring people whose search results would reflect poorly on them. [55]

Women interviewed by Dr Jane report – with good reason – that their economic circumstances will be further compromised by the fact that once reputationally damaging material is circulated about them online, it is all but impossible to remove and has a potentially unlimited lifespan, thereby potentially sabotaging their work prospects indefinitely. The blurring of personal and professional contexts, meanwhile, means that work-related abuse spills into women's personal lives, and vice versa [34].

The Australian writer and broadcaster Ginger Gorman – herself a victim of extreme trolling resulting from her work as a journalist – points out that while many businesses have social media policies prohibiting staff from bringing their employers into disrepute (that is, policies to mitigate harm to organisations), a double standard is at play because employers rarely demonstrate similar concern for employees in the form of policies to protect staff members from the harm that might come to them online in the course of using social media to perform their work duties [35]. In her investigative journalism around the issue, Gorman identifies the problem as having serious implications for occupational health and safety (OHS) acts, regulations, and codes of practice, as well as potentially being “a sleeping giant in terms of negative social impact” [35].

Gendered cyberhate with economic dimensions can involve lost productivity, missed work opportunities, being blamed for attacks, and/or having upsetting and harmful experiences trivialised by colleagues or employers. Women interviewed by Dr Jane describe agitating about whether to change careers or to perform existing work differently, with some women deciding to leave particular jobs – and even particular lines of work – completely. Subjects repeatedly describe finding themselves in an oppressive double bind in that retreating from the internet means they are unable to perform the tasks required to do their jobs, yet staying online and enduring abuse and harassment can also hinder their productivity. This is because of the time required to block, delete, report, and engage in damage control, as well as because of the potentially disabling psychological fall-out. The latter not only have the potential to adversely impact women’s ability to perform their daily work tasks, but are likely to interfere with career advancement going forwards. Even those female workers who experience milder but ongoing cyber harassment are likely to be suffering harm given research showing that less “intense” but more frequent harmful workplace experiences are just as detrimental to women’s occupational well-being as single instances of more extreme sexual coercion and unwanted sexual attention [36].

Much of the cyberhate women receive at work involves abuse and harassment which represent contraventions of laws and policies in many nations, as well as of various international labour treaties, conventions, and recommendations. For instance, it breaches many sections of the International Labour Organization (ILO) guidelines with regard to States’ and employers’ obligations *vis-à-vis* women worker’s rights and gender equality [37]. Unfortunately, there exists a vast gulf between the best practice ideals advanced by bodies such as the ILO and the reality of working life for many women. This gulf yawns particularly wide for those women whose experience of workplace abuse and harassment occurs primarily via electronic means, and within “precarious” labour contexts. One reason for this is because many of the laws and institutional provisions that emerged to provide (albeit limited) protection to women from sexual harassment in the Fordist workplace provide little to no protection to women in new economy sectors such as games design [38].

A further complication concerns the fact that antagonists who attack women in non-work-related contexts (for instance after a relationship break-up or in response to a targets’ political activism) are able to exploit the idiosyncrasies and reach of networked communications technology to inflict far-reaching professional and financial damage to women in ways which do not map neatly onto extant paradigms *vis à vis* laws and workplace policies. We might

expect such acts of economic vandalism to be covered via civil or criminal legislation, such as via laws relating to defamation, intimidation, stalking, extortion, threatening death or bodily harm, and so on. Yet, as we explain in Section 4.2. of this submission, law enforcement agencies and the courts face an uphill battle in terms of responding to online forms of abuse [39]. Anecdotal accounts of police advising female cyberhate targets to withdraw from the internet in response to being threatened and abused suggests that at least some authorities do not appreciate the realities of contemporary working life with regard to the centrality of the cybersphere [5: 88-92].

To illustrate those dimensions of cyberhate that can be understood as involving a new form of workplace harassment and/or economic vandalism, we offer three, additional Australian examples.

Kath Read: Kath Read is a Brisbane-based librarian and body acceptance activist who uses multiple social media platforms to campaign against fat shaming. Her high online visibility has resulted in a flood of “fat, ugly, bitch” messages that can arrive daily in their hundreds. These have included threats to decapitate her with a chainsaw, and to smash her face in with a hammer if she is spotted in the street. Over eight years of abuse, Kath’s detractors have extended their efforts offline, on one occasion leaving a note in her mailbox reading, “Hi fat bitch, I see this is where you live”. (When Kath reported this note to police, a male officer told her to, “get offline and stop being so confident”.) Groups of strangers have collaborated to sabotage Kath’s paid freelance projects, as well as her primary job as a librarian. For example, a hate web site exclusively devoted to Kath names her workplace and suggests people pay her “a visit” there. A doxing forum has also published her work details, leading to her being signed up for multiple mailing lists and requests for contact with weight-loss clinics, gyms, personal trainers, diabetes specialists, heart clinics, and bariatric surgeons. As a result, she has received numerous phone calls via her workplace’s call centre number. Removing herself from these databases took weeks, and required some mortifying conversations with her employer. In addition to having already lost freelance work, Kath is concerned that the hate-driven content available about her online might deter potential employers [34].

Amy Grey: Since commencing writing commentary for The Guardian in 2013, Amy Gray has regularly received abusive and threatening messages. These have included, “You are just another bipolar whore who sluts and sucks dick for cash and free drinks ugly pig”, and – in response to a column about global terrorism – messages calling her a “fucking Arab lover” who would be beheaded and raped with a knife. Amy says the latter frightened and “fucked [her] up ... badly”. In addition to being upset, however, she is angry that she must endure abuse simply for having a job, and “not even a particularly well-paying” job at that. Cyberspace is

her workplace, she says, and she feels like she's being hounded out of it, and can no longer talk or express herself there in the way she once could [34].

Tracey Spicer: The Australian newsreader, documentary maker, journalist and writer Tracey Spicer has received what she calls online "sexualised violence" on an almost daily basis since 2013. This has included mob attacks by men's rights groups from around the world and threats to rape and murder Spicer and her children. A particularly savage attack unfolded in April 2014 when a travel column she had written resulted in:

really vile attacks from all over the world . . . "you deserve to be raped but you're too ugly. I wouldn't want to fuck your children anyway", this kind of just horrific stuff . . . I got quite scared . . . I was actually frightened to return to my own home . . . because the threats were so horrific – not only to me, but to my children. It gave me an awful fright . . . some of the ones that were from the men's rights groups in America were quite explicit . . . "we know that you're in Australia, but we know where your kids go to school", this kind of stuff . . . there were absolutely direct threats to kill me and to rape me and to kill the children.

While I'm not the kind of person who's prone to anxiety . . . I did go through months where I was more housebound. I didn't want to go outside that much. I looked over my shoulder while I was outside. . . . When they pile upon you, you feel like you don't want to go out the front door because there might be a mob out there. That's how you feel. It almost manifests physically in your mind that there are people out there with pitchforks.

Spicer decided against reporting the attacks to police because she was worried that this would result in more and worse attacks on both herself and her children. Instead, she attempts to "mute" or "block" anyone who attempts to engage her by using violent language, as well as changing both the content and style of her writing:

I'm a little bit ashamed to admit this, but I've been frightened to write too many full-on columns ever since then . . . I've really eased off on writing edgy columns because it scared the shit out of me. [31]

3.5. *New digital divide*

The internet is strongly linked with prosperity and career progression. Access to broadband is recognized as playing a vital economic and social role in all nations, with digitally disadvantaged workers facing barriers to full economic participation that their more digitally advantaged peers do not [40]. For instance, a multitude of economic advantages accrue to those who are able to use the internet continuously at work and at home, and who are skilled at "curating their professional self-presentations" on social media [41: 574-5]. Research into the impact of gendered cyberhate in workplace contexts shows that while women might *seem* to

have full and unfettered access to the internet, in practice, the hate and harassment they experience might be severely constraining their ability to use it. Moreover, those women who most depend on unrestricted access to the internet and social media platforms to earn their living might be particularly prone to receiving cyberhate. As such, the cumulative disadvantages of gendered cyberhate should be understood as constituting an emerging, economic dimension of existing, gender-related digital divides [42].¹² Further, this is a digital divide that is insidious in that it involves barriers to equity and full participation online that are not as easy to identify and measure as those barriers relating to access to computer hardware and network connections.

¹² “Digital divide” is a term used to discuss online equity, and refers to differences between population groups in terms of access and of information and communications technologies.

4. RESPONSES TO CONSULTATION QUESTIONS

Having explained in Section 2.5. why we do not offer responses to the last five consultation questions posed in the AHRC's HRT Issues Paper, and having also explained in Section 3 why we share the AHRC's view that cyberhate is a critically important issue, in this section we now turn to the task of presenting our responses to the AHRC's first five consultation questions.

4.1. RESPONSE TO QUESTION ONE

What types of technology raise particular human rights concerns? Which human rights are particularly implicated?

4.1.1. The technologies at issue

In the context of this submission, the types of technology giving rise to particular human rights concerns range from hardware, software (aka “apps”), and platforms¹³ including websites, emails, and multiplayer games (i.e. technologies associated with Web 1.0 era¹⁴ iterations of the internet), as well as Wi-Fi, broadband internet access, wireless computing, high resolution web cams, mobile devices such as smart phones and tablets, and social networking sites such as Facebook, Twitter, Instagram, and Tumblr (i.e. technologies associated with the Web 2.0 era) [66]. New surveillance technologies are also increasingly being misused to stalk, intimidate, harass, humiliate, and coerce intimate partners, particularly girls and women. This includes: using electronic means to remove access to targets’ bank account funds; blocking emails and phone calls from friends and family members; installing GPS trackers on targets’ vehicles; and circulating false and/or intimate information about targets online [67].

4.1.2. The human rights at issue

This section draws on examples discussed throughout this submission to demonstrate how cyberhate – which is critically reliant on the above technologies – infringes on at least ten of the Articles defined in the Universal Declaration of Human Rights (UDHR).

Article 7 of the UDHR grants everyone a right to **equal protection under the law**. However, this protection is currently not being extended to victims of cyberhate. Not only are the interests that cyberhate compromises not given due recognition, but there are systemic reasons why it remains difficult to even get recognition for the importance of these interests. Furthermore, the people whose interests cyberhate threatens and compromises are especially vulnerable – jwomen, children, LGBTQI+ people, as well as people from cultural and ethnic minority groups. At present, these vulnerable groups are not afforded an equal right to protection under the law.

¹³ The term “platform” here is used to describe a combination of hardware and software.

¹⁴ See the top of Section 3.1. for a brief discussion of “Web 1.0”.

Article 8 of the UDHR describes everyone’s right to **justice in a court or tribunal** if their rights are violated. However, when the importance of the interests compromised by cyberhate is not even recognised, when the technology involved makes it effectively impossible to collect evidence, when police and other authorities do not even have a framework for recognising what evidence should be collected and what comprises credible evidence, and when jurisdictional ambiguities mean that there is often no court or tribunal that will hear the grievances of these victims, this right cannot possibly be given substantive protection.

Article 12 of the UDHR states: “No one shall be subjected to arbitrary interference with his **privacy**, family, home or correspondence, nor to attacks upon his honour and **reputation**. Everyone has the right to the **protection of the law** against such interference or attacks.” [2] However, extremely intrusive interferences with privacy, and personally and economically devastating attacks on reputation, are the *lingua franca* of cyberhate. For instance, recall the practice of “doxing” that we described in Section 3.1., which involves the publishing of personally identifying information, usually to incite internet antagonists to hunt targets in “real” life [5: 34–5]. Similarly, the example from Section 3.2. of Jebidiah James Stripe, the 28-year-old American man who impersonated his former female partner on the internet site Craigslist and posted, alongside invitations for men to show up for rough sex, his ex-partner’s street address, which resulted in another man showing up, binding and blindfolding her, and then raping her at knifepoint, is another clear case in point. In the latter case, not only was the target’s right to privacy breached, but the breach also had horrendous personal consequences [16: 5-6]. For evidence of how cyberhate infringes the right to be free of arbitrary interference with one’s reputation, reflect on the examples cited in Section 3.4., which demonstrate the ways in which online abuse has the power to destroy women’s reputations, and the permanent effects this can have on their livelihoods and future employment prospects. Findings from the Pew Research Center show that of those people targeted for physical threats and sustained harassment, about a third feel their reputations have been damaged [21: 7]. The legal scholar Danielle Keats Citron’s work shows that schools have fired teachers whose naked photos have appeared on revenge porn sites, while a government agency terminated a woman’s employment after a co-worker circulated her nude photograph to colleagues [55]. Cogent, too, is the fact that most employers rely on candidates’ online reputation to filter applicants. Finally, as per our comments above in regards to Articles 7 and 8 of the UDHR, victims of cyberhate are patently not afforded equal protection of the law against such interferences and attacks.

Article 13 of the UDHR grants everyone the right to **freedom of movement** within their country, to leave their country and enter another, and to return home from abroad. The examples provided throughout Section 3.1. provide a plethora of disturbing material to demonstrate just how severely cyberhate infringes on targets' freedom of movement. Recall, for instance, the Amnesty International UK study published in November 2017 which showed among other things that more than a quarter of the women who were targets of cyberhate (27 per cent) received direct or indirect threats of physical or sexual violence, and that one-third (36 per cent) felt their physical safety had been threatened [14]. Leaving home when one feels afraid for one's safety clearly satisfies this criterion. Alternatively, consider the examples of Zoë Quinn, Anita Sarkeesian, or Brianna Wu, all of who received such credible threats of extreme violence that all three women fled their homes, while Sarkeesian fearing for her life cancelled her speaking event at Utah State University. In one sense, such victims still retain the right to freedom of movement, but they choose to not exercise it. In another sense – and clearly the more relevant and important one – such victims' substantive ability to exercise their right of freedom of movement is severely compromised by cyberhate, either by being frightened into not leaving their homes, or leaving their homes, or cancelling travel plans.

Article 19 of the UDHR grants everyone the right to **freedom of expression**. As things stand, however, at present victims of cyberhate are prevented from exercising this right, while perpetrators of cyberhate defend their abusive actions by claiming that they are merely exercising their right to freedom of expression, and people who should know better end up pandering to the perpetrators rather than stepping in to protect the victims. The examples offered in Section 3.4. of how women attract a vastly disproportionate amount of cyberhate by comparison to men in the field of journalism. And the two examples cited in Section 3.2., of Annie Nolan whose benign comments related to her twins attracted a deluge of vitriol and physical threats, and Waleed Aly whose sole reason for attracting the wrath of cyberhate offenders is his cultural background and media platform, again show that cyberhate has extremely damaging effects on people's freedom of expression. The right to freedom of expression is clearly not protected when abuse is permitted to masquerade itself as an exercise of the right to freedom of expression, and when vulnerable populations are silenced through the very same intimidation and abuse that masquerades itself as a benign and protected exercise of freedom of expression.

Article 21 of the UDHR states that “[t]he will of the people shall be the basis of the authority of government” [2], and it grants everyone an equal right to **take part in governance** of their

country. With governments as well as businesses increasingly shifting their operations online and, in particular, with governments looking to online spaces and using the digital footprints left by their citizens to ascertain who those citizens are and what is of importance to them, what matters is not only having access to the internet in a *thin sense*,¹⁵ but more importantly having access to the internet in a *thick sense* — namely, not having impediments to doing such things as participating in online forums where information and conversations increasingly take place, where friendships are made and broken, where jobs are found, where new interests are developed, and where education (including access to the media) increasingly happens. For many people, it no longer makes sense to talk about them doing things online, simply because the word “online” has become otiose. We do not do our shopping online, or pay our bills online, or engage in conversations online – we just do our shopping, pay our bills, and engage in conversations. The centrality of the online world to people’s daily lives is constantly on the rise, and not only is this the channel through which citizens learn about what their governments are doing and get an opportunity to express their views in response to government calls for submissions, but rather people’s behaviour online also leaves traces of what people care about and what they do. These traces, as much as direct interaction with government, are increasingly used to inform government decisions. However, for victims of cyberhate who experience so much harassment, intimidation, and abuse – as well as threats, some genuine, others bluff, but both equally threatening – that they either reduce the amount of online behaviour, or constrain the forms of behaviour online (for instance by not expressing their views for fear of being yet again the targets of cyberhate), or who just take their lives offline, their views are never even heard, nor are they able to actively participate in governance. As we noted in Section 3.5., cyberhate is a significant factor which contributes to the digital divide, and those who withdraw from online life for fear of persecution and no recourse to remedies or protections, they patently do not get an opportunity to take part in governance of their country, and this is another clear way in which cyberhate violates one of the most important human rights for the just and healthy operation of a democratic state.

Article 23 of the UDHR grants everyone, among other things, the **right to work** and to **good working conditions**. To see how cyberhate infringes on this right, reflect upon the study findings noted in Section 3.1. which mentioned that “[i]ncreasing numbers of women are also

¹⁵ In this context, the “thin sense” of “having access to the internet” means such things as, for instance, having access to an ADSL, NBN, or cellular broadband connection at home or nearby.

reporting instances of: cyberstalking; rape blackmail videos; large groups attacking individuals (sometimes with the explicit purpose of causing job loss or career derailment).” Alternatively, recall how the cyberhate that Zoë Quinn endured resulted in employers withdrawing job pending offers to her boyfriend, how Anita Sarkeesian was forced to cancel a speaking event at Utah State University after an anonymous emailer threatened “the deadliest school shooting in American history” if her talk went ahead as planned [46], and Brianna Wu’s observation that the threats made against her safety led her to comment that “[e]very woman I know in the industry is scared. Many have thought about quitting” [44]. The examples cited in Section 3.4. of the online abuse women receive in workplace environments are simply too numerous to mention. As the final example, consider our discussion in Section 3.5. about the impact of cyberhate on creating a new and deeply damaging digital divide. Given the findings of research into the impact of gendered cyberhate in workplace contexts, which shows that while women might *seem* to have full and unfettered access to the internet, in practice, the hate and harassment they experience might be severely constraining their ability to use it – coupled with the fact that unfettered access to social media results in multiple economic and career benefits – the cumulative disadvantages of cyberhate should be understood as constituting an emerging, economic dimension of existing, gender-related digital divides. Cyberhate clearly infringes on the right to work and to good working conditions.

Article 27 of the UDHR protects everyone’s right to **participate in the cultural life** of their community, and to **enjoy the benefits that this confers** including access to arts and sharing in the benefits of scientific advancements [2]. As we noted in the previous paragraph, in relation to Article 21 of the UHDR, life increasingly takes place in online spaces. To gain access to goods and services – let alone to even learn about their existence – one needs the freedom to develop an online presence and the skills to find them. Moreover, though, the internet has created a new space in which culture exists. For instance, transgender support groups exist solely online, or jointly online and offline, and this has enabled a group of people who once (before the advent of the internet) felt isolated and lacking in support to realise that they are not alone, and that their views and interests and needs matter. The range of social media platforms on which a large part of the population interacts by sharing stories, pictures, videos, and so forth has created digital cultures which – quite apart from gaining access to anything in the physical world – comprise a cultural domain that people should not be prevented from inhabiting. To participate in the cultural life of your community, and to enjoy the benefits that this confers, requires that impediments be removed, especially when those impediments are

created by perpetrators who inflict unjust harms on vulnerable minorities. Until the problem of cyberhate is taken seriously, and until these vulnerable groups' interests are protected so that they are not effectively disbarred from participating in online communities due to threats that are played down and ignored, these vulnerable groups' ability to exercise this right will continue to be violated.

Article 28 of the UDHR states: “Everyone is entitled to a **social and international order in which the rights and freedoms** set forth in this Declaration **can be fully realized.**” [2] In the preceding paragraphs of this section we briefly recounted numerous examples to demonstrate the variety of ways in which cyberhate thwarts its victims' ability to exercise their rights. Both social and technological factors combine to create this state of affairs. This point is especially important because the social and international order is not a given fact about the world – it is not a law of nature, or something like the weather that we can do little about – but rather it is something that is in our power to shape. What this entails is that the duty to create the social and international order in which the rights and freedoms of cyberhate victims can be fully realized falls squarely on our collective shoulders, but in particular on those in power who can do something about it. We elaborate on this point in the next paragraph.

Lastly, Article 29 of the UDHR is special because it defines not just *rights*, but rather also the *duties* that correlate with – or undergird – those rights, as well as clearly spelling out *whose* shoulders those duties fall upon. Here's how the Simplified Version of the UDHR put these two points regarding duties: “**We all have a responsibility to** the people around us and should **protect their rights** and freedoms.” [1] For comparison, here is how the original wording of the UDHR stated this same point: “Everyone has duties to the community in which alone the free and full development of his personality is possible” [2]. We cite these passages to emphasise the critical role that the UDHR attributes to society as the duty-bearer for protecting human rights, and the reason why this is important for our present argument is because of the multitude of ways in which, at present, various organisations and individuals in society are failing to exercise this critical duty that underpins the rights defined by the UDHR.

As we shall argue in Sections 4.2., 4.3., and 4.4. of this submission – i.e. in the context of our answers to the AHRC's second, third, and fourth questions – there are numerous things that could be done to protect people from cyberhate, but yet these things are either never or rarely done.

For instance, the scholarly community and the media do the victims of cybercrime a great disservice – and if the wording of Article 29 is taken literally, an injustice not just a disservice – by using vague, generic, and sterilised words like “trolling” and “offensive behaviour” to describe the harmful conduct of cyberhate perpetrators, and similarly vague, generic, and sterilised words like “upset” and “offence” to describe the impact that cyberhate has on the victims.¹⁶ While this language may be easier on the eye, as well as compliant with the etiquette of polite conversation, the effect of persisting with us of such inoffensive terminology is that the true face of cyberhate is never put on display and consequently either difficult to recognise for what it is or simply not recognised at all. In the academy, scholars and teachers alike have a duty to investigate cyberhate and to develop terminology and examples that will accurately portray this harmful and rights-curtailling online behaviour, whether in talks at conferences and symposia, in academic publications, or in the coursework that they present to their students. The media’s reach to the public clearly also imposes on it a duty to accurately portray cyberhate, its harms and costs, and its victims, in all of the finely textured details, rather than using words and examples that will not cause offence to viewers. The police, legislators, and judges, also have a duty to find out what cyberhate actually is, whom it impacts, and how it impacts them, for otherwise they become complicit in perpetuating the impression that it is not as deserving of our attention as other forms of cyber abuse and those that have already been defined in legislation as cybercrimes. They will also be blameworthy for many important omissions including failure to develop clear guidance about what evidence victims need to present when reporting cyberhate, whom they should present this evidence to, and what standards must be met to ensure that evidence is valid, as well as legislating or using other regulatory mechanisms such as financial incentives to ensure that designers of technology create safer online spaces and methods for obtaining the requisite evidence, as well as laws that protect these victims’ rights.

Higher education institutions are also remiss in not being quick enough on the uptake to ensure that ethics is integrated into the curricula of science and technology degrees, and for their failure to impart their students with knowledge and skills to apply design methods such as Value-Sensitive Design (VSD), Default Choice Architectures (DCA), and Socially Responsible Innovation (SRI),¹⁷ to ensure that when these students enter industry they will

¹⁶ For a lengthy discussion, please see Section 4.2.1. below.

¹⁷ Please see Section 4.3.1. for a discussion of these methods.

design online environments that meet the same standards of safety as those we expect of the people who design, create, and maintain the physical environment. Technology designers and manufacturers, which includes scientists and engineers, as well as platform operators – e.g. Facebook, Twitter, Instagram, YouTube (a subsidiary of Google), etc – are remiss in their duty to apply methods like VSD, DCA, and SRI in their own work, since it results in the creation of technologies and online spaces that are unsafe and provide no way of capturing evidence of and reporting abuse, and create safe havens for perpetrators of cyberhate by continuing to provide people with the ability to create effectively endless anonymous and untraceable accounts that are disposable,¹⁸ that impose no costs on those who use them for abuse, and that do not prevent the spread of (or provide feasible means to withdraw from circulation) personally and economically damaging information once it has been posted and propagated through the internet.

Beyond citing examples, though, our point is that Article 29 of the UDHR places the duty that correlates to the rights defined in the UHDR – that is, the duty to ensure that the rights of those around us are protected – on everyone’s shoulders. In the modern interconnected world, the phrase “those around us” necessarily includes the people who live in other countries, since that is the reach of the online spaces in which significant harms are inflicted on a great many victims. At present, we are therefore remiss in not divesting ourselves of our duties to protect those around us from cyberhate. Likewise, they too are remiss in not divesting themselves of their duties to protect us from cyberhate. In an interconnected world, where people living in different countries can both be victims or perpetrators of cyberhate, we all owe each other duties to protect one another. Given the scope of this duty, however, and the way it includes natural and artificial persons from around the globe, we believe that a human rights approach, together with its institutions and well-recognized as respected as well as tested mechanisms of an international human rights framework, are needed to coordinate this effort.

4.1.3. Concluding comments

In the above discussion we have provided merely a few examples (taken from lengthier discussions elsewhere in this submission) of how a recognition of the importance of the interests that cyberhate threatens and compromises, would require a concomitant recognition that even under the UDHR cyberhate is clearly a human rights issue; that it is an issue which

¹⁸ For further discussion, please see Section 4.2.8.

involves the range of technologies that underpin a range of internet services listed in Section 4.1.1. above; and that it affects a great number of different people. Given how the international human rights framework has evolved ever since the original formulation of the UDHR in 1945 by the United Nations, including the International Covenant on Civil and Political Rights (ICCPR) [3], and the International Covenant on Economic, Social and Cultural Rights (ICESCR) [4], our discussion should clearly be taken only as the *start* of this conversation. Cyberhate threatens and compromises human rights in numerous ways, and it is critical that this be given due recognition and that steps be taken to protect these human rights.

4.2. RESPONSE TO QUESTION TWO

Noting that particular groups within the Australian community can experience new technology differently, what are the key issues regarding new technologies for these groups of people (such as children and young people; older people; women and girls; LGBTI people; people of culturally and linguistically diverse backgrounds; Aboriginal and Torres Strait Islander peoples)?

In Section 3 above we provided numerous examples of the highly damaging impact of cyberhate on its targets, and we also argued that women are disproportionately impacted due to the nature of the cyberhate that is aimed at them. Nevertheless, there are two important reasons – both of which are based on insights we obtained from the Cyberhate Symposium – why in responding to this second consultation question we shall not proceed by merely reciting the litany of issues that we already discussed in Section 3 above.

Firstly, as we noted in Section 2.2.2. above, our analysis of the points that were raised at the Cyberhate Symposium revealed three distinct categories of issues: (i) challenges to recognising, reporting, and tackling cyberhate; (ii) potential solutions to cyberhate; and (iii) twelve stakeholder groups who are involved in the prior two categories. However, what point (i) clearly entails is that the issues which cyberhate raises actually need to be divided into *two* distinct groups. First, there are the harmful impacts that cyberhate has on a range of interests that human rights protect. Second, there are also numerous factors that make cyberhate into an extremely challenging problem to tackle. Now, clearly our aim in putting together this submission is to tackle the first group of issues – i.e. to protect people’s rights from being violated by cyberhate. However, given that the second group of issues obscures and thus makes the first group of issues extremely difficult to tackle, what our response to the AHRC’s second question actually needs to do, is to elucidate this second group of issues as clearly as possible, rather than merely reiterating the first group of issues. After all, until these issues are clearly identified, brought to the AHRC’s attention, and then addressed, cyberhate will continue to be an intractable problem.

Going one step further, we believe that the very reason why in some circles cyberhate continues to be overlooked or played down as “just words” and/or “just the internet”, is precisely because of how this second group of issues obscures the problem. If this is right – and in this section we shall cite compelling evidence for the fact that it is right – then it is highly likely that the problems that cyberhate causes and the groups on which it impacts are in fact significantly more numerous than what we described in Section 3.

The likelihood that at present all we are seeing is the tip of the iceberg as far as the problem of cyberhate is concerned, feeds into the second reason why we do not wish to proceed in this section merely by reciting the litany of issues which we already discussed in Section 3 above. Namely, as the participants at the Cyberhate Symposium noted on multiple occasions, it is critically important to investigate – rather than assume that we already know – the intersectional issues that cyberhate raises. Although the particular manifestation of attacks on women indeed inflicts extremely serious consequences on their targets, this neither entails that all women (e.g. regardless of age, ethnicity, socio-economic status, location, or occupation) necessarily experience the same kinds of cyberhate attacks, or that they experience them in precisely the same ways, nor does it entail that other groups are immune from similarly devastating cyberhate attacks. Although the Cyberhate Symposium participants noted that the ways in which women were attacked and the consequences of those attacks were clearly in a class of their own, they also urged that other groups including children, older people, people from cultural and linguistic minority groups, people in the LGBTQI+ community, as well as people with physical and mental disabilities, are just as likely to be targets and to suffer comparably severe consequences. The majority of attacks discussed in Section 3 above are ones committed against women. However, at the same time, the category “victims” is not homogenous, and for this reason it is critically important to recognise how this intersectionality might result in potentially many different manifestations of cyberhate, in different reasons why those manifestations adversely impact important interests, and why these instances of cyberhate may require different responses (e.g. from police, from courts, or other institutions) in order to protect the different groups of victims.

Thus, drawing on discussions from the Cyberhate Symposium, as well as other supplementary sources, in this section we shall aim to bring to light the wide range of insidious structural, institutional, and technological factors – as well as interactions between them – that make cyberhate into a problem that is difficult to notice let alone to understand. In a nutshell, our case will be as follows. Firstly, setting aside our stated aim to quote unexpurgated examples of cyberhate in Section 3 above, in many walks of life people still lack the language and examples in which to adequately describe cyberhate. Secondly, this makes it difficult for people and institutions to recognise cyberhate, to report it, to legislate about it, and to study it. Thirdly, it is unclear to whom evidence of cyberhate can even be reported. Fourthly, and finally, a range of technological factors compound the aforementioned difficulties — in particular,

technology creates special difficulties for victims *vis à vis* reporting cyberhate, as well as difficulties for authorities *vis à vis* investigating it.

4.2.1. *Inadequate language and examples*

The language that is used to report cyberhate is often extremely nebulous, generic, and bland. For example, in media narratives, it is often referred to via the vague, catch-all term “trolling” (a word associated with pranking in internet sub-cultures), which further compounds the difficulties involved in having cyberhate recognised, let alone recognised as a seriously inappropriate form of behaviour. In scholarship, too, it has traditionally been referred to via generic descriptors such as behaviour which is “hostile”, “graphic”, “in bad taste”, and so on, and/or illustrated via only the most tame and expurgated examples.¹⁹ However, what such euphemisms, and censored and linguistically pasteurised examples fail to capture is the violence and threatening nature of contemporary cyberhate.

To get a sense of why we so emphatically emphasise the need to describe and discuss cyberhate by using realistic language and examples, compare the difference between the following:

- a. Women are receiving sexually explicit rape threats online.
- b. Women are receiving sexually explicit rape threats online such as, “I will fuck your ass to death you filthy fucking whore. Your only worth on this planet is as a warm hole to stick my cock in”.²⁰

Real-life examples of cyberhate such as these – and the others offered in the previous section²¹ – are likely unpleasant and unsettling to read. Our concern, however, is that academic squeamishness about citing unexpurgated examples of cyberhate has the potential not only to misrepresent the phenomena but to stifle – in a very tangible way – research into this critical issue. For example, one of the authors of this submission received feedback from a peer reviewer chastising her for citing unexpurgated examples of gendered cyberhate in a grant application. She was advised that even researchers “examining avant-garde sexual practices

¹⁹ For a lengthy discussion of this topic, please see [6].

²⁰ This was tweeted at the feminist writer Sady Doyle [12], and is an example drawn from and discussed at greater length in [5: 13-14].

²¹ It bears emphasis that the examples we provide in this submission should only be treated as offering a glimpse of the disturbing character and wide diversity of the material to which the term “cyberhate” refers. For more examples, please see [23], [62], or [63].

don't use such language" and was asked to remove it [5: 103]. Given that this man went on to question her overall argument on the grounds that he himself had not noticed any misogynistic hate speech online, it was frustrating that he did not wish to be exposed to the very evidence that would have supported her thesis. Insight into the limits of scholarly thinking on this issue is also evident in another reviewer's comment – provided to one of the authors of this submission under similar circumstances – that cyberhate can easily be ignored and dismissed because it is usually poor writing containing multiple spelling and grammatical errors, and therefore lacks credibility [6: 72]. These positions – which we contend are indicative of a larger scholarly phenomenon²² – suggest that cyberhate should be ignored because it is somehow both (i) too unsavoury to discuss in civil discourse, and yet (ii) too innocuous to warrant serious consideration. We reject both these framings which together are clearly incoherent.

In a nut shell, our case is that rich, nuanced, and finely-textured *language*, as well as detailed, realistic, and evocative *examples*, are essential to convey and to recognise the diverse texture of (i) implicated behaviours, (ii) the harms caused, (iii) the range of victims, and (iv) why these things are serious not innocuous. Without adequate language and examples, it is difficult not only to identify and describe instances of cyberhate, but to recognise why it is a serious problem.

4.2.2. No standards

Another factor that contributes to the reasons cyberhate can go unrecognised – by victims, onlookers, and potentially even offenders – is the lack of clear and recognisable standards of what qualifies as acceptable online conduct. Admittedly, paradigmatic examples of what constitutes good – or at least acceptable – conduct in the physical world differ from place to place and change over time. However, in most of the ways that count, there is little disagreement about what standards apply in offline contexts. For instance, shouting in someone's face, or threatening to rape them, perhaps using sexually suggestive but intentionally ambiguous language, are not regarded as acceptable standards of conduct. In the physical world, examples of what qualifies as due process – that is, how to proceed when things go wrong – are also available. For instance, victims of physical and sexual assault – or, equally, witnesses – can report to the police, who follow procedures for interviewing victims, witnesses, and collecting evidence, as well as relatively clear and tested laws that guide police in their

²² For a lengthy discussion of this topic, please see [6].

investigations. However, this is distinctly not the case in relation to the online world, where standards of good (or minimally acceptable) conduct in different online contexts are often extremely unclear.

As we shall now go on to explain, these three factors – (i) the vague, generic, and bland language, (ii) an almost complete lack of examples to characterise the different forms that this behaviour takes, and its serious impact on victims, and (iii) the lack of standards to codify and provide exemplars of acceptable and unacceptable online conduct – have a range of very important ramifications which contribute to the reasons cyberhate is ignored, played down, or simply not even recognised.

4.2.3. Barriers to recognition of cyberhate: effects on offenders, victims, and witnesses

It is one thing to be a target of or a witness to cyberhate – or even to be a person who inflicts cyberhate on someone else – but it is quite another to *recognise* particular kinds of behaviour as instances of cyberhate, and to understand why that behaviour is objectionable. With ambiguous and generic language, no examples, and no clear and recognised standards to consult, some people may unfortunately fail to comprehend the gravity of their behaviour and the damage it inflicts on their targets. This would not justify or excuse what they do, but it might partly explain why they do it. Furthermore, these same three factors may also prevent victims and witnesses from recognising the problematic behaviour as instances of something that is unacceptable. Their uncertainty about what they are experiencing or observing may even lead them to question themselves. For instance, victims may doubt that they are entitled to demand the offender to stop, and observers may doubt whether they should step in and either render assistance or offer to act as witnesses (because ambiguities tend to lead people to vacillate). Thus, the first reason the three factors discussed above are deeply toxic is because, on the one hand, they raise the chances that some people may inflict cyberhate on others without even realising it and because, on the other hand, they create very unhelpful epistemic barriers to recognising cyberhate.

4.2.4. Not knowing what evidence to collect

These three factors also means that victims and witnesses have little guidance about what evidence to collect in order to substantiate that they were targeted by – or that they witnessed cyberhate. In part, this is simply because – without a clear idea of what cyberhate is, or that it is something objectionable – a person will not know what features of an interaction they need to record to collect evidence that the interaction involved was, indeed, cyberhate. The other

reason victims and witnesses may not know precisely what evidence to collect, has more to do with the clarity that standards of evidence could provide. For instance, are screenshots with time-stamps and information about what computers or mobile devices the various parties were using – and perhaps at what street addresses or geographic locations they were using their respective computers or mobile devices – adequate forms of evidence? Is it enough for such screenshots to only show the usernames of the parties involved in the interactions, or do they also need to show the details of the profiles associated with those usernames, the IP addresses of the computers or mobile devices from which those parties were connected, and maybe even other data too? And if other data is required, then precisely what might that be and how might it be obtained if one is not an IT expert?

The present point may sound simple, but we and our Cyberhate Symposium participants all believe it is critically important. The problems created by these three factors have a profound impact on people’s ability to identify whether and in what ways their behaviour, how they are treated, or how they see another person being treated, is inappropriate. This, in turn has downstream effects for victims’ and witnesses’ ability to determine what evidence to collect to substantiate the claim that what happened was, indeed, an instance of cyberhate.

4.2.5. Where should cyberhate targets or witnesses even take their evidence?

Unfortunately the situation is significantly worse than what we have so far described, because victims or witnesses of cyberhate cannot currently expect to obtain a serious hearing or adequate response even if they do present at a police station bearing all the evidence they can muster. Multiple anecdotal accounts suggest that – instead of attempting to apprehend perpetrators – police have been known to chastise female cyberhate victims for some aspect of their mode of conduct on the internet and to tell them that the onus is on them to alter their behaviour online or somehow “take a break” from the internet altogether²³ – a phenomenon with deeply unsettling parallels to the victim-blaming that remains so prevalent in offline instances of sexual and other violence against women. These anecdotal accounts of police ignorance, insensitivity, and inaction are supported by empirical data. According to the World Wide Web Foundation, in 74 per cent of Web Index²⁴ countries (including many high-income nations), law enforcement agencies and the courts are failing to take appropriate action in

²³ For a lengthy discussion of this issue, please see [5: 88-92].

²⁴ The World Wide Web Foundation’s Web Index covers 86 countries and measures the web’s contribution to social, economic and political progress.

response to acts of gender-based violence online, while one in five female internet users live in countries where harassment and abuse of women online is extremely unlikely to be punished [73: 15, 4].

Indeed, the present situation is so deeply troubling that the group that focused on investigating issues surrounding activism at our Cyberhate Symposium proposed that what is needed is a National Day of Reporting. Specifically, they proposed that women – as well as girls, LGBTIQ+ people, people from ethnic, cultural, and linguistic minority groups, and anyone else who has been exposed to cyberhate – stage a protest (akin to those associated with the SlutWalk [71] or Women’s March [72] movements) by gathering up screenshots of all the online abuse, harassment, and threats they have experienced, and then forming orderly queues at their local police stations to report it. It was noted that, even if only a fraction of the people who currently endure this toxic abuse show up at police stations to present their evidence, these queues would potentially snake for kilometres. As Jessamy Gleeson – producer for the feminist talk show *Cherchez La Femme*, organiser for the Melbourne chapter of SlutWalk and a Cyberhate Symposium participant – acknowledges, officers behind the front desks of police stations that day would be totally overwhelmed. However, she adds that, “jamming the gears of the system by taking this problem to the streets would be the point. Not to gratuitously cause trouble, but to highlight the broader problem that our police simply lack adequate resources to deal with this escalating problem.” We concur with Gleeson that if the institutional and technological impediments to reporting cyberhate are not addressed – impediments that account for the reasons cyberhate is currently ignored, played down, and not recognised – then this leaves little other option but to engage in activism [64].

To appreciate the institutional and technological impediments to reporting cyberhate that motivated Gleeson – and the activist group at the Cyberhate Symposium – to propose the need for such a National Day of Reporting, consider the range of institutional and structural impediments to reporting cyberhate. These are impediments that prevent society from even starting to collect in an official capacity the examples, as well as prevalence and impact statistics, about cyberhate that together constitute the critical first steps required to formulating remedies to the problem.

4.2.6. Effects on police

Relatedly, the lack of explicit guidance on such matters as where the law stands in regards to these different kinds of behaviour, whether such behaviour falls into the jurisdiction of the

police or some other authority, precisely how police should handle reports of cyberhate, or what evidence they need to gather, create significant institutional impediments to reports of cyberhate being filed and followed up on by police. When the face of crime starts changing from offline to online offences, as is happening right now, police need explicit new guidelines (and new training on how to put those guidelines into practice) about such things as what questions to ask cyberhate victims and witnesses in order to obtain relevant and useful information and evidence. This is not just to make sure that police do not make mistakes – that due to a lack of clarity about what is and is not problematic behaviour, they treat some people too harshly and others too leniently – but because following procedures that were once useful for investigating offline offences, is unlikely to result in asking the right kinds of questions that yield useful information and evidence. Given the changing face of crime, and the need to gather data on the nature and impact of cyberhate, these issues are in urgent need of attention.

As we briefly noted above, police are critically reliant on the existence of clear legislation, and information about whether the legislation has been tested or not. However, for the same reasons as those that impact on perpetrators, victims, witnesses, and police, so, too, legislators need access to the right information – i.e. that rich, nuanced, and finely-textured language, and the detailed, realistic, and evocative examples, as well as authoritative standards about what counts (and for what reasons) as inappropriate behaviour online – in order to even create the requisite offences in the first place. But without specific offences and information about whether convictions have been made when perpetrators are charged with those offences, police cannot charge individuals who engage in the problematic behaviours.

While Section 4.2.5. above might seem critical of police, we wish to highlight the degree to which police are incredibly under-resourced, and that, to serve their community, a range of institutional factors must be urgently addressed. The police representatives attending the Cyberhate Symposium made it clear that members of law enforcement are becoming increasingly aware of – and frustrated by – the inadequate resources at their disposal to deal with the problem, and they are eager to find ways to help.

4.2.7. Legal impediments to reporting cyberhate

Another factor that creates institutional barriers to cyberhate being reported, is that online offences of all sorts create steep conceptual, doctrinal, and pragmatic problems for the law. This part of our discussion raises deep and important issues that highlight the urgent need to approach cyberhate from an international human rights perspective. This was the point at which

we gestured towards the end of Section 4.1.2. above, because it clearly highlights why a human rights approach is critically needed in any effort to tackle cyberhate.

Unlike conventional criminal offences which typically occur in a specific physical location, whenever people interact on the internet, at a purely conceptual level it is incredibly difficult to say *precisely where* their interactions unfold. [74] After all, the victim may be in one country, the offender in another, and their interactions may take place on third-party servers located on yet another country's soil. Providers of fora can also host their operations with one company (in one country) one day, and move them to another hosting company (in another country) on the following day. For all international internet-mediated interactions, servers in many locations around the world need to collaborate with one another to do such things as look up and translate domain names into IP addresses, or to progressively move a message along from server to server until it reaches its destination. This feature of online interactions – that there simply is no clear place where such interactions occur – creates distinct challenges with flow-on consequences for policing and enforcement, as well as prosecution of offenders.

What it is critical to notice is that the question of *where* cyberhate occurs is not even one that yields to empirical investigation, since it concerns a conceptual issue. Namely, when actions and interactions take place in a virtual place, is it even legitimate to point to any given physical location and say, "That is where those (inter)actions took place?" Plausibly, the answer is "no." A similar conceptual issue is encountered when we ask where a telephone conversation between two people on opposite sides of the Earth occurred. In one country? In the other? Or on the wires and satellites that carry the digital signals that encode their voices? Another example might be when a person located in one country sends a letter containing a pathogen like anthrax to someone in another country. Was the crime committed in the location from which the letter was posted, or at the destination where it was received? Perhaps the right answer is that the telephone conversation and anthrax attack happened on Earth, or even – to broaden the scope so that the physical location of telecommunication satellites is included – within the Earth's orbit, and thus to say something of a similar sort about where instances of cyberhate take place.

But although in one sense such an account would be correct – cyberhate and other cyber abuses and cyber-offences indeed happen within the border of the Earth's orbit – this answer creates a very unhelpful legal situation. Namely, that offences require a jurisdiction in which they are defined, which is usually co-extensive with a physical place with physical borders. If we cannot name a concrete physical place, then we will not be able to pinpoint a specific

jurisdiction either, with the upshot that there will be no jurisdiction in which the offence can even be reported let alone tried. Given that crimes are defined in jurisdictions, unless specific legislation is created – or unless enforceable international agreements are put into effect and then followed – to recognise cyberhate as a criminal offence internationally, then instances of cyberhate that involve international interaction may not even fall into any jurisdiction.

On a more practical note, without adequate international inter-jurisdictional agreements and cooperation – for instance, about which nation state, province, or governing body will prosecute cyber-offences of various kinds – there may be no one to whom such offences can even be reported, let alone through which they could be investigated, pursued, and prosecuted.

4.2.8. Technological barriers to reporting cyberhate and related factors

In this section we discuss features of the technologies involved that further compound the difficulties we already discussed above in regards to reporting cyberhate.

Anonymity and disposable accounts. Firstly, it is incredibly easy to create effectively anonymous accounts – that is, without the need to provide any personally-identifying information – and to enter misleading details. Although in some cases, user profiles also show such details as how long an account has existed, and sometimes even information about the account’s reputation as rated by other online users on the platform, this is by no means a common practice. This makes it extremely easy and inexpensive – often, effectively free – for those who wish to do so to create an endless stream of disposable accounts that, once used to engage in cyberhate and other forms of cyber abuse, to dispose of those accounts with no ill consequences, and no way of tracing the account back to the real person who created and used it for the purpose of abuse. Similarly, various features of the internet such as proxy servers, virtual networks, and other technologies also make it effectively impossible to pin down either the physical location of the computer or device on which an offender’s account was located, or the physical location from which they accessed it, let alone to identify the specific person involved. Given this setup, victims of cyberhate can hardly even be expected to collect and report evidence about their abuser.

Impermanence of digital data. Another problem with expecting victims of cyberhate to collect evidence about the abuse to which they have been exposed, and thus with the expectation that people will even be in a position to report such evidence to authorities, is the ease with which electronic data can often be permanently erased. A related problem stems from the fact that the pace at which interactions can take place on internet forums can either be too

fast for victims to even get a chance to collect the evidence, or so slow that by the time the abusive nature of the interaction has unfolded, too much history has scrolled by without a trace in order to still collect sufficient evidence. Given that interactions are not things that happen in an instant, but something that unfold over some duration of time – sometimes beginning in a seemingly innocuous way, and only later unfolding in a way that makes it evident that abuse is taking place, by which stage it may already be too late to start recording *all* the evidence – just how much history of the interaction (and with what kinds and degree of detail) needs to be recorded?

A characteristic feature of the digital world is that, unlike the physical world in which people leave traces of their behaviour – traces that can be gathered and inspected using forensic techniques – the digital world is not specifically designed to create digital traces, and often no data can be left behind to testify to what events actually unfolded.²⁵ The upshot this has is that unless a platform on which people interact happens to implement a permanent log of interactions that cannot be altered by users – for instance, that prevents users from deleting their own abusive posts – then victims who attempt to collect evidence may simply be thwarted if that data was erased before they had a chance to collect evidence.

Mechanisms to report cyberhate and other cyber-offences. Although some platforms implement mechanisms through which users can report other users for unacceptable behaviour, this is by no means a universal feature of all platforms, nor is there evidence to suggest complaints are currently likely to result in any effective remedies.²⁶

Police lack the tools and training to investigate reports of cyberhate. If what is needed for society to start taking cyberhate seriously is for official reports to be lodged with the authorities, then what this requires is for reporting mechanisms to lodge the requisite evidence not just with

²⁵ We are cognizant of the fact that there appears to be a tension between, on the one hand, saying that once sensitive data is posted on the internet it can be impossible to ever remove it while, on the other hand, also saying that in the digital domain information lacks permanence. It sounds as if we are trying to have our cake and eat it too. However, our point here is not a conceptual one but an empirical one, and it relates to the fact that data stored in memory does not in general terms leave traces once the memory is filled with new data. Exceptions to this rule exist – for instance, that data can often be retrieved from a corrupted hard drive, and that technologies like blockchain have been specifically designed to record – and to keep track of changes to all data. However, these exceptions are not the rule, and they do not impact on the point which we are making at present.

²⁶ For a lengthy discussion of this issue, please see [5: 95-97].

platform operators, but directly with police. In part this requirement stems from the fact that otherwise the data could be tampered with, and in part because – given the impermanence of data and the speed at which interactions can take place – time is of the essence. However, even if such mechanisms that reported cases of cyberhate directly to police were implemented, not only would this likely result in a flood of reports which the police would be unable to process fast enough, but police – just like other people – lack the tools and the training to investigate the cases properly. Furthermore, given that digital information is (at least in principle) infinitely malleable/modifiable, and modifications of digital content will often leave little or no trace of when, how, or by whom the data was modified, this would raise the further challenge for police of how to validate reports if those reports were lodged by victims themselves, in order to ensure that the evidence could be certified as valid, rather than leaving open the possibility that it might have been tampered with. The problems that we discussed above under the headings “Anonymity and disposable accounts” and “Impermanence of data” are just as difficult to overcome for police as they are for other users. These problems are, after all, effects of characteristic features of the technology, and nobody currently has the means to address these problems.

4.2.9. Tip of the iceberg

When the reasons are systematically investigated and properly spelled out, it is hardly a challenge to explain why cyberhate remains an under-appreciated and under-recognised problem. We lack the language and examples in which to describe the problem, which in turn makes it difficult for people and institutions to recognise it, report it, legislate about it, and study it. The technological barriers to reporting cyberhate only compound these problems, and the fact that even if victims had the means to get around these issues and obtain the needed evidence, it would still be far from clear to whom such evidence should even be presented, and what the authorities could even do about it, completes the picture. This complex mix of factors – which includes multiple stakeholders, institutions, technologies, and norms – explains why cyberhate is often not reported, and why without official reports being lodged, the victims and impacts keep being overlooked and played down as “just words” and/or “just the internet”. It also underscores the need to be mindful to engage in the sort of intersectional research that we discussed in the introductory comments to Section 4.2..

4.3. RESPONSE TO QUESTION THREE

How should Australian law protect human rights in the development, use and application of new technologies? In particular:

- a) What gaps, if any, are there in this area of Australian law?
- b) What can we learn about the need for regulating new technologies, and the options for doing so, from international human rights law and the experiences of other countries?
- c) What principles should guide regulation in this area?

As we argued at the top of Section 4.2., cyberhate creates two distinct groups of issues that need to be addressed. First, there are the harmful impacts that were discussed in Section 3. Second, there are the factors that make it difficult to even start noticing, let alone addressing, cyberhate's damaging impacts. In this section, our responses to the AHRC's question 3 will aim to address both groups of issues, by listing a number of potential solutions which involve legislative and regulatory responses.

4.3.1. Regulations to encourage or mandate hardware and software developers, including Platform operators, to employ methods such as Value-Sensitive Design (VSD), Default Choice Architectures (DCA), and Socially Responsible Innovation (SRI)

Our first recommendation for how Australian law could protect human rights in the development, use, and application of new technologies in order to start tackling cyberhate involves the proposal to develop new legislation and regulations that encourage or mandate hardware and software manufacturers, as well as platform operators, to employ methods such as VSD, DCA, and SRI.

In a nutshell, these three design methods can be described as follows.

VSD involves treating values – for instance, safety, equality, and sustainability, but equally values like the importance of not permitting complete anonymity in cyberspace – as no less important vis à vis design considerations than technical requirements such as how much power a device is allowed to consume, or how much heat it is allowed to dissipate. An example of VSD is the way that some toilets are designed in such a way that there are no level surfaces on which a person can temporarily put down, and later forget to pick up, and important item like a wallet, a passport, or their telephone. Although a sign saying “Do not forget your valuables as you leave” could in theory achieve the same result, in practice people do not read signs, whereas if there is no (e.g.) toilet roll holder on which to place your wallet, mobile phone, or

passport, then you will have to put it in a safe place instead where you will not forget it by accident.

DCA involves the development of systems in such a way that by default, the easiest way to use the systems is in a way that will be pro-social rather than anti-social. For instance, the organ donation policy according to which by default everyone opts in as an organ donor unless they explicitly state that they wish to opt out, sets up “opt in” as the default option, with the result that countries that have adopted this policy do not have shortages of organ donors, while countries that adopt the opposite policy do have organ donor shortages. Most importantly, because DCA leaves people the option to do something other than the default option, it does not undermine anyone’s liberty since it leaves it up to them whether they wish to do something other than the default option.

Finally, SRI involves the engagement of designers with stakeholders at an early stage in the development of an artefact or a system of one sort or another in order to ensure that the needs and values of the stakeholders are properly taken into account. Another feature of SRI is to also attempt to predict potential pitfalls of technology ahead of time, in order to avoid people getting hurt and then needing to blame, when the whole situation could have been avoided in the first place by doing better testing, which involves attempting to predict how people might use different technologies and what effects that may have down the road.

The reason why VSD, DCA, and SRI are so important is because as we noted in Section 4.2.8. above noted, fundamental features of the implicated technologies play a critical role in why cyberhate is such a difficult problem to address at present. Just like mobile phone SIM cards cannot be purchased or registered without first presenting adequate forms of ID, so too social media platforms could require people to provide adequate forms of ID to register for an account. This does not need to mean that everyone would immediately have access to everyone else’s personal information, but only that this information would be safely stored so that it could be retrieved in case an account is used for cyber abuse and the perpetrator cannot be identified. Likewise, technologies like blockchain make it impossible to alter data without leaving a trace of who modified it and when they modified it, which could create a supportive technological mechanism to ensure that evidence is not contaminated, tampered with, or destroyed. Technology is not something that just comes into existence, but rather we make it, and thus if these features are so problematic that they impinge on people’s human rights, then the onus is on technology designers to implement technology with better features.

Stated bluntly, technology needs an ethical upgrade. Another option might be to consider a complete ban on instant/disposable accounts, and/or accompanied by the slow unlocking of full account functionality on various platforms once the user has earned it by proving that they are good netizens. This would not only give others greater protection, but it would also give them a vested interest in not treating that account as disposable, since they would have invested time and effort into opening up its various features. While so-called "real name" policies are open to criticism, we think it is at least worth considering the advantages of new account applicants having to provide enough evidence of who they are as real, flesh-and-blood humans, as per our earlier suggestion. Details that could then be used by authorities to track down offenders, regardless of whether they abandon their accounts after committing abuse.

The use of VSD, DCA, and SRI applies to platform operators as well, who should clearly start considering the safety and other implications of the platforms they create. Hardware, software, and platform designers must take responsibility for designing safer spaces — just like the safety we build into offline environments. For the very same reasons why we do not design streets and walkways with dark and dingy nooks and crannies where innocent passers-by can be cornered and attacked, and why office buildings are these days designed in such a way that visitors can have a safe path of exit in case their host (whoever occupies that office normally) turns out to be abusive, so too online environments as well as the software and hardware that they are built on top of should have such safety designed into them. Why should safety standards only apply to streets and buildings, when online highways and platforms are increasingly becoming the places we spend our time?

A clear advantage of all of these approaches is that they are proactive/promotive rather than reactive/protective. Instead of waiting for people to first start getting hurt, and only then thinking about doing something about it — and often the first step is pointing the blaming finger rather than actually doing something to stop the problem from recurring — they instead involve taking active steps to make sure that people will not get hurt. Naturally, even with the best of efforts, things will still go wrong, either because of oversights, or hardware and software bugs, or because people figure out how to use the well-designed technology for nefarious purposes. However, until we actually start designing technology in this way, we simply cannot expect to start seeing fewer problems. And if the problems that need to be tackled require that technology receives an ethical upgrade, then in our view the Australian Government needs to legislate and/or create regulations that either mandate or incentivise technology designers and companies to take cyberhate interests seriously into account. A reactive/protective component

should still have a role to play, in part because of the just-mentioned fact that things will still go wrong even if everyone does their best effort to design technology that supports the stakeholders needs. However, in addition, if human rights would need to be violated in order to design, implement, or use technology that is safe for cyberhate victims (of for any other stakeholder), then clearly human rights should provide a constraint on what things designers are allowed to do in order to achieve otherwise socially valuable outcomes.

However, given current practices, unless the Australian Government acts to create the necessary legislation and/or regulations to mandate or incentivise technology manufacturers to employ these techniques, then it is unlikely that this will occur. Hence our proposal that this is something that requires the Australian Government to take action.

4.3.2. Regulations to encourage or mandate higher education institutions to incorporate ethics, VSD, DCA, and SRI subjects into their courses

The same considerations as the ones we noted above, are also why we recommend that the Australian Government should make it a requirement that ethics, VSD, DCA, and SRI subjects are taught to engineering and design students. Not as a soft subject or “politically correct” inconveniences, but as serious core components which imbue students with the know-how to enable them to “baking in” ethical and not just practical functionality into software, hardware, and platforms. In fact, our case is that learning to build ethical functionality into artefacts and environments should be an integral part of the training of every designer and engineer, and just as important as learning to build and program any other functional requirement.

Given that the students of today are the people who will build the IT systems of tomorrow, it’s never too early to start educating the next generation of people who will build the technological infrastructure of our society so that they know how to design it in a way that protects people — from cyberhate as well as from other technologically-induced or technologically-enabled problems.

4.3.3. Create new legislation that imposes liability, fines – and potentially criminal charges – for failing to take adequate measures to create unsafe technologies

Because legislation and regulation that has no bite for failing to observe its requirements is likely to be ignored, we also recommend that the Australian Government should create legislation that imposes liability, fines, and potentially even criminal charges on those who fail to take adequate measures to create safe technologies — i.e. those who cannot show records of

having employed VSD, DCA, and SRI techniques in the process of developing, implementing, testing, and deploying their wares.

Where a duty of care is imposed on someone to take certain measures, then a failure to carry out that duty without an adequate justification would constitute an act of negligence – negligence by omission, in this case – and if the negligent omission results in another party’s harm, then at a minimum this should provide a ground for the party to sue the technology producer in negligence. However, whether the mechanisms involve private litigation, fines, or criminal charges, these are details that would need to be worked out base on a range of different factors. Our point is simply that to ensure that people are protected from cyberhate, there need to be ramifications for those who fail to take adequate care in the process of developing and deploying their technologies.

4.3.4. Reward universities that include ethics, VSD, DCA, and SRI subjects in their courses

By similar reasoning to the above, we also recommend that there needs to be a mechanism to incentivise universities that include ethics, VSD, DCA, and SRI subjects in their courses. Because of the critical role that universities play, and the fact that they are already underfunded, we do not favour using punitive measures to discipline those universities which do not comply. Furthermore, because students who have received an education that includes these courses are more likely to create technology that reduces harm, the cost savings in terms of harm reduction could be used to fund the system of rewards for including ethics, VSD, DCA, and SRI subjects in their courses.

4.3.5. Amend legislation regarding the publication of ostensibly offensive material

To tackle the problems we discussed in Section 4.2.1. above, regarding the lack of language and examples, it may also be beneficial to consider amending legislation that governs the publication of ostensibly offensive material. In order for cyberhate to be recognised, examples of it need to be published in unexpurgated form, and made available to the public, rather than dressing cyberhate in the clothes of civility, which is what results when vague, generic, and bland language is used to describe cyberhate. Furthermore, unofficial barriers to publishing reporting on cyberhate in its original form should also be considered. For instance, if the prospect as a consequence of publishing such material, a publisher may be sued by a reader who finds such material offensive, then this too would discourage the publication of such material and thus educating the public.

To be clear, we do not mean to endorse legislation that permits the publication of intentionally offensive material. Rather, what we are proposing is that sensitively crafted legislation is needed to ensure that such material can even be published in the first place. This also does not mean that gory headlines should dominate the front pages of newspapers, but only that disincentives to publishing such material may play an important role in keeping the public uninformed about what cyberhate is and why it has such detrimental impacts.

4.3.6. Set up a taskforce to develop standards of acceptable online conduct

As per our discussion in Section 4.2.2., in order for a number of institutions as well as individuals to have the necessary resources to refer to in order to tell whether particular conduct is acceptable or unacceptable – in this case, to be able to distinguish cyberhate from other things – a taskforce needs to be set up to develop standards of acceptable and unacceptable online conduct. As with any standard, the aim is not to provide a document or manual that will answer every question with utter certainty, but rather to provide a common reference guide to codify in public way what conduct meets, and what conduct fails to meet, the standards of acceptable online behaviour.

Because what this involves is ultimately the public's standards of decency, and this evolves over time, stakeholder engagement would be crucial in this process, as would be the establishment of an ongoing process or institution/authority that is charged with maintaining this document. Likewise, because interactions these days often take place between people in different places around the world, and standards clearly differ from place to place as well as changing over time, whichever authority is charged with this task should cooperate on an international level to develop multilateral agreements.

Once the above institution is established, it should also be charged with the task of developing an action plan for how to address the range of problems that we detailed in Sections 3 and 4.1., as well as the authority to investigate complaints received from people who allege that they have been victims of cyberhate and other cyber abuses. This may require the development of standards for what constitutes appropriate evidence, but rather than treating this as a task to be accomplished in full completion before the authority begins to investigate any complaints, we instead propose that the right way to proceed is through a casuistic (that is, a case-by-case) manner in which engagement with the cases progressively leads to the development of the needed standards. This institution should also oversee the process of potentially developing new offences under which cyber offenders can be charged for cyber

abuses such as cyberhate. Such a move would also create a welcome reference point for legislators, for police, and for platforms to consult, as well as for designers of software, hardware, and platforms, when they set out to design technology that takes publicly-endorsed and institutionally-verified values into account.

4.3.7. Set up a task-force to determine how best to enable police to protect the community, while the other matters are being handled

As we explained in Section 4.2.6. above, the police are currently extremely under-resourced to handle reports of cyberhate. The problems in many ways are not dissimilar to those that affect individuals and other institutions — namely, inadequate language and examples, no standards, no certainty about what laws do or don't apply, inadequate training about how to proceed with online rather than traditional offences, no specialist tools to carry out their investigations, and no training on how they would use such tools even if they did exist. However, the police are likely to be the first port of call — they are the ones to whom victims and witnesses are likely to turn to, if they need to report an incident to the authorities. For this reason we believe that while the other matters discussed in this section are being handled, the police should be provided with the right advice on how to proceed in the meantime.

However, because we believe that the problems involved require a proactive/promotive design solution that includes the VSD, DCA, and SRI components, as well as an approach that tackles the legal jurisdictional issues at an international level (see Section 4.3.8. below), we do not believe that the best strategy is to focus on arming the police with currently available tools to tackle currently existing problems. In fact, there is good reason to believe that it would even not be the best use of resources to allocate significant pools of funding to the development of future-technology-based tools, given that these would ultimately tackle tackle problems that exist because of faults with the way the technology is currently being designed (i.e. problems which should be tackled by designing better technology), and problems with the other issues we have identified (for which we have suggested solutions in this section).

We fully believe that the police should be supported so that they can protect the community and enforce the law. However, we do have reservations about whether the best way to support the police in this regard is by developing tools, training, and providing other resources, when the problems require a structural solution with elements of the sort that we have been describing in this section.

4.3.8. Cooperate with international organisations on developing an international approach to resolving jurisdictional issues

One of the biggest challenges that we identified are the jurisdictional lacunae created by the fact that the traditional physical borders of nation states, provinces, and localities, cannot be expected to serve as a valid basis for policing online behaviour. And while our principal recommendation is that instead of developing legislative responses to cyberhate, it is better to focus on developing proactive/promotive design-based responses – i.e. ones that employ VSD, DCA, and SRI, as per our comments in Section 4.3.1. above – it is still nevertheless critically important to ensure that in those cases where proactive/promotive strategies designed to protect people from cyberhate and other cyber abuses fail, there are still reactive/protective strategies to fall back upon in order to protect human rights.

Given that this requires collaboration between nation states on a legislative level, and that the focus here should be on the protection of human rights, we believe that the AHRC can play a critical role in leading the development of an international legal framework that clarifies the jurisdictional problems that currently lead to victims often having no one to turn to when their human rights are violated, because the jurisdictional issues that apply to online matters have not been sorted out.

4.4. RESPONSE TO QUESTION FOUR

In addition to legislation, how should the Australian Government, the private sector and others protect and promote human rights in the development of new technology?

In this section we list a number of non-legislative ways in which the Australian Government, the private sector, and others could help protect the human rights of cyberhate victims.

Firstly, in line with our recommendation in Section 4.3.1. above, technology designers and manufacturers as well as platform operators should employ VSD, DCA, and SRI in the development, testing, and deployment of their technologies.

Secondly, in line with our recommendations in Section 4.3.2. above, higher education institutions should include subjects on ethics, VSD, DCA, and SRI, within their courses.

Thirdly, in line with our recommendation in Section 4.3.5. above, the media should report on cyberhate and other cyber abuses and cyber offences using the appropriate language and unexpurgated examples.

Fourthly, the Australian Government should allocate funding to set up and operate the taskforces as per our recommendations in Sections 4.3.6. and 4.3.7. above.

Fifthly, given that VSD, DCA, and SRI provide the best approach to tackling cyberhate, and harm reduction will result in cost savings for Australia, the Australian Government should promote research into- and use of these methodologies within higher education institutions and in industry. This could take the form of competitive Australian Research Council grants for projects judged likely to result in the development of more ethically appropriate technologies.

Sixthly, researchers at higher education institutions who study cyberhate and other cyber abuses, should report their findings by using rich language and quoting unexpurgated examples, rather than perpetuating the politeness, or squeamishness about reporting offensive language and images accurately, since doing this only serves to make the phenomena they study and write about sound mysterious, as well as unproblematic.

Seventh, and in significantly more detail, platform operators need to develop a better income model than the current one based on advertising revenues. The income model that platforms like Facebook, Twitter, and Instagram use is fundamentally at odds with protecting online users' interests. Advertising-generated income increases with clicks — the more people click on a story, tweet, or picture, the more people are exposed to advertising, which platform operators display for a fee which they charge to their clients who wished to have their

advertisement displayed. Unfortunately, the effect this has is that it incentivises the posting, retention, and promotion of sensationalist content, and content that emphasises conflicts and rage, since that is precisely the kind of content that attracts the most clicks. However, given this income model, there is little incentive for platform operators to screen for instances of posts that will generate outrage or upset, since those are precisely the sorts of posts that will attract attention.

To make our point, consider this example. In its 2018 documentary “Inside Facebook”, for instance, the Australian Broadcasting Corporation current affairs program Four Corners aired footage filmed by a reporter training undercover as Facebook moderator [68]. It depicts the reporter being shown video footage of an adult man kicking, beating, and stamping on a small boy and being told it was an example of the sort of Facebook post moderators should mark as “disturbing” rather than being deleted. This was despite the fact that Facebook had been receiving complaints about the post from child abuse campaigners since 2012. In the same documentary, Roger McNamee, a former mentor to Facebook founder Mark Zuckerberg, described such content as the “crack cocaine” of Facebook:

It’s the really extreme, really dangerous form of content that ... attracts the most highly-engaged people on the platform. If you’re going to have an advertising-based business, you need them to see the ads, so you want them to spend more time on the site. And what Facebook has learned is that the people on the extremes are the really valuable ones, because one person on either extreme can often provoke 50 or 100 other people and so they want as much extreme content as they can get. [69]

Competition between platform operators and success are based on click-ability and the potential to prompt outrage, which gives the platform operators no incentive to screen and intervene, and may even scare away users to migrate to other platforms, which will in turn harm their profits. At a structural level, there is currently nothing in place either to punish a platform’s bad conduct, or to reward its good conduct. Finally, although in theory community moderators could perform some of the role of policing such online spaces, the platforms provide no incentives to attract public moderators who might do this. In effect, anyone who chooses to do this ends up performing a thankless task, often exposing themselves to the risk of reprisal from those who feel aggrieved that their behaviour was targeted by the moderator as stepping out of line, and the moderator effectively performs free work for the platform’s benefit, even though the platform loses profits if the moderator does a good job.

4.5. RESPONSE TO QUESTION FIVE

How well are human rights protected and promoted in AI-informed decision making? In particular, what are some practical examples of how AI-informed decision making can protect or threaten human rights?

Our comments in this section are very brief. In effect, an interim measure, while the proposals we discuss in Sections 4.3. and 4.4. above are developed, could be to develop Artificial Intelligence based algorithms that attempt to monitor platforms, email, and other electronic channels through which people may be exposed to cyberhate, and to filter out (or at least alert human operators about) offending content.

However, although this idea may have some promise, and it might warrant being investigated, the reason we fear that it may ultimately not turn out to be effective is that since even humans are susceptible to failing to recognise such things as humour and irony, and since context – such as the context that a history of interactions within a long term friendship might provide – critically informs whether specific utterances are offensive or not, it is difficult to imagine how AI, which is after all trained on human-generated judgments, could meet this challenge.

Again, this is not intended as a rejection of this idea, but simply as a statement of the steep challenge that implementing an AI-based cyberhate detection system would involve.

5. REFERENCES

- [1] Universal Declaration of Human Rights (UDHR), Simplified Version. Civics and Citizenship Education website sponsored by the Australian Government Department of Education, Employment and Workplace Relations, and maintained by Education Services Australia. Accessed from https://www.civicsandcitizenship.edu.au/verve/_resources/FQ2_Simplified_Version_Dec.pdf on 4 October 2018.
- [2] Universal Declaration of Human Rights, GA Res 217A (10 December 1948). Accessed from <http://www.un.org/en/universal-declaration-human-rights/> on 4 October 2018.
- [3] International Covenant on Civil and Political Rights, opened for signature 16 December 1966, 999 UNTS 171, (entered into force 23 March 1976).
- [4] International Covenant on Economic, Social and Cultural Rights, opened for signature 16 December 1966, 993 UNTS 3 (entered into force 3 January 1976).
- [5] Jane, Emma A. (2017a), *Misogyny Online: A Short (and Brutish) History*. LA, London, New Delhi: SAGE.
- [6] Jane, Emma A. (2015), "Flaming? What flaming? The pitfalls and potentials of researching online hostility", *Ethics and Information Technology* 17(1): 65–87.
- [7] Nolan, Annie, personal communication with Emma A. Jane on 31 July 2015.
- [8] Willis, Charlotte (2015), "Project co-host Waleed Aly: 'Why I'm passionate about not being on social media'", [news.com.au](http://www.news.com.au/entertainment/tv/project-cohost-waleed-aly-why-im-passionate-about-not-being-on-social-media/news-story/f90d2ff7a1b7fbd50fdac822c48741e2), September 24. Accessed from <http://www.news.com.au/entertainment/tv/project-cohost-waleed-aly-why-im-passionate-about-not-being-on-social-media/news-story/f90d2ff7a1b7fbd50fdac822c48741e2> on 7 November 2018.
- [9] Lattouf, Antoinette, personal communication with Emma A. Jane on 27 May 2016.
- [10] @BELIMBLA4 (2015), Twitter, May 30. Accessed from <https://twitter.com/belimbla4/status/604379648055742464> on 23 August 2016.
- [11] crackshot (2015), "Waleed Aly blames non muslims for paris attacks", *liveleak.com*. November 17. Accessed from http://www.liveleak.com/view?i=d4e_1447758255&comments=1 on 23 August 2016.

- [12] Doyle, Sady (2011a), “But how do you know it’s sexist? The #MenCallMeThings round-up”, Tiger Beatdown, 10 November. Accessed from <http://tigerbeatdown.com/2011/11/10/but-how-do-you-know-its-sexist-the-mencallmethings-round-up/> on 7 November 2018.
- [13] “Cyber violence against women and girls: A world-wide wake-up call” (2015), The United Nations Broadband Commission for Digital Development Working Group on Broadband and Gender. Accessed from http://www.unwomen.org/~media/headquarters/attachments/sections/library/publications/2015/cyber_violence_gender%20report.pdf on 7 November 2018.
- [14] “More than a quarter of UK women experiencing online abuse and harassment receive threats of physical or sexual assault - new research” (2017), Amnesty International UK, 20 November. Accessed from <https://www.amnesty.org.uk/press-releases/more-quarter-uk-women-experiencing-online-abuse-and-harassment-receive-threats> on 7 November 2018.
- [15] Tran, Cindy and McNab, Heater (2015), “I never meant to hurt a single soul”: Mother of twins who became an internet sensation when she posted hilarious signs about her frustrations hits back at angry critics, Daily Mail Australia, 13 July. Accessed from <http://www.dailymail.co.uk/news/article-3159153/This-photo-staged-did-not-actually-walk-like-Young-mum-hilarious-signs-twins-hold-pram-hits-online-critics.html#ixzz4I1DErLYJ> on 7 November 2018.
- [16] Citron, Danielle Keats (2014a), *Hate Crimes in Cyberspace*. Cambridge, Massachusetts, and London, England: Harvard University Press.
- [17] Powell, Anastasia and Henry, Nicola (2015), “Digital Harassment and Abuse of Adult Australians: A Summary Report”, Tech & Me Project, RMIT University. Accessed from: https://research.techandme.com.au/wp-content/uploads/REPORT_AustraliansExperiencesofDigitalHarassmentandAbuse.pdf on 8 November 2018.
- [18] Gardiner, Becky, Mansfield, Mahana, Anderson, Ian, Holder, Josh, Louter, Daan, and Ulmanu, Monica (2016), “The dark side of Guardian comments”, The Guardian, 12 April. Accessed from <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments> on 8 November 2018.

- [19] “Australian media still a Blokesworld in 2016” (2016), Media, Entertainment and Arts Alliance, 6 March. Accessed from <https://www.meaa.org/news/australian-media-still-a-blokesworld-in-2016/> on 8 November 2018.
- [20] “Cyberhate Symposium” (2017), Parliament of New South Wales, 8 October. Accessed from <https://www.parliament.nsw.gov.au/Hansard/Pages/HansardResult.aspx#/docid/HANSARD-1820781676-74084/link/101> on 8 November 2018.
- [21] Duggan, Maeve (2014), “Online Harassment”, Pew Research Center, 22 October. Accessed from <http://www.pewinternet.org/2014/10/22/online-harassment/> on 8 November 2018.
- [22] Neary, Ben (2010), “2nd man gets 60 years in Wyo. Internet rape case”, Ventura County Star, 29 June. Accessed from <http://www.vcstar.com/news/2nd-man-gets-60-years-in-wyo-internet-rape-case-ep-368408277-348997991.html> on 8 November 2018.
- [23] Jane, Emma A. (2014), “‘Back to the kitchen, cunt’: speaking the unspeakable about online misogyny”, *Continuum: Journal of Media & Cultural Studies* 28(4): 558-570. DOI: 10.1080/10304312.2014.924479, p. 563.
- [24] Doyle, Sady (2011b), “The girl’s guide to staying safe online”, *In These Times*, 17 November. Accessed from https://www.inthesetimes.com/article/12311/the_girls_guide_to_staying_safe_online/ on 8 November 2018.
- [25] Elam, Paul (2011). “Stalking Sady Doyle”, *A Voice for Men*, 18 November. Accessed from <http://www.avoicemen.com/feminism/feminist-lies-feminism/stalking-sady-doyle/> on 8 November 2018.
- [26] Sandoval, Greg (2013), “The End of Kindness: Weev and the Cult of the Angry Young Man”, *The Verge*, 12 September. Accessed from <http://www.theverge.com/2013/9/12/4693710/the-end-of-kindness-weev-and-the-cult-of-the-angry-young-man> on 8 November 2018.
- [27] Sarkeesian, Anita (2015a), “Talking publicly about harassment generates more harassment”, *Feminist Frequency*, 29 October. Accessed from <https://feministfrequency.com/2015/10/29/talking-publicly-about-harassment-generates-more-harassment/> on 8 November 2018.

- [28] Sarkeesian, Anita (2012b), “Image based harassment and visual misogyny”, *Feminist Frequency*, 1 July. Accessed from <http://feministfrequency.com/2012/07/01/image-based-harassment-and-visual-misogyny/> on 8 November 2018.
- [29] Smith, Matthew (2018), “Four in ten female millennials have been sent an unsolicited penis photo”, YouGov UK, 8 September. Accessed from <https://yougov.co.uk/news/2018/02/16/four-ten-female-millennials-been-sent-dick-pic/> on 8 November 2018.
- [30] Mantilla, Karla (2015), *Gender trolling: How Misogyny Went Viral*. Santa Barbara, CA: Praeger.
- [31] Jane, Emma A. (2017b), “Feminist Fight and Flight Responses to Gendered Cyberhate”, in Segrave, Marie, and Vitis, Laura (eds.), *Gender, Technology and Violence*. London and New York: Routledge.
- [32] Day, Elizabeth (2013), “Caroline Criado-Perez: ‘I don’t know if I had a kind of breakdown’”, *The Guardian*, 8 December. Accessed from <https://www.theguardian.com/society/2013/dec/08/caroline-criado-perez-jane-austen-review-2013> on 8 November 2018.
- [33] Henry, Nicola and Powell, Anastasia (2015), “Embodied harms: Gender, shame, and technology-facilitated sexual violence”, *Violence Against Women*: 21(6): 758–799. DOI: 10.1177/1077801215576581, p. 765.
- [34] Jane, Emma A. (2018), “Gendered Cyberhate as Workplace Harassment and Economic Vandalism”, *Feminist Media Studies*, special edition on Online Misogyny. DOI: 10.1080/14680777.2018.1447344.
- [35] Gorman, Ginger (2015), “Why aren’t employers offering their staff social media self-defense training?”, *Daily Life*, March 15. Accessed from <http://www.dailylife.com.au/news-and-views/dl-opinion/why-arent-employers-offering-their-staff-social-media-selfdefense-training-20150313-1436k3.html> on November 8 2019.
- [36] Sojo, Victor. E., Wood, Robert. E. and Genat. Anne. E. (2016), “Harmful Workplace Experiences and Women’s Occupational Well-Being: A Meta-Analysis”, *Psychology of Women Quarterly* 40(1): 10-40.

- [37] “ABC of women workers’ rights and gender equality (second edition)” (2007), International Labour Organization. Accessed from http://www.ilo.org/wcmsp5/groups/public/---dgreports/---gender/documents/publication/wcms_087314.pdf on 8 November 2018.
- [38] Elliot, Amanda (2015), “Gamergate: Gender at work in the new economy” (seminar), School of Social and Political Sciences, The University of Sydney, 3 August.
- [39] “Web Index: Report 2014-15” (n.d.), Web Index. Accessed from http://thewebindex.org/wp-content/uploads/2014/12/Web_Index_24pp_November2014.pdf on 8 November 2018.
- [40] “The state of broadband 2015” (2015), The Broadband Commission for Digital Development, United Nations Educational, Scientific and Cultural Organization, September. Accessed from <http://www.broadbandcommission.org/documents/reports/bb-annualreport2015.pdf> on 8 November 2018.
- [41] Robinson, Laura., Cotten, Sheila. R., Ono, Hiroshi, Quan-Haase, Anabel, Mesch, Gustavo, Chen, Wenhong., Schulz, Jeremy, Hale, Timothy M., and Stern, Michael J. (2015), “Digital inequalities and why they matter”, *Information, Communication & Society*: 18 (5): 569-582.
- [42] Jane, Emma A. (2017), “Gendered cyberhate: a new digital divide?”, in M. Ragnedda and G. W. Muschert (eds.), *Theorizing Digital Divides*. Oxon: Routledge.
- [43] Jason, Zachary (2015), “Game of fear”, *Boston Magazine*, May. Accessed from <http://www.bostonmagazine.com/news/article/2015/04/28/gamergate/> on 8 November 2018.
- [44] Stuart, Keith (2014a), “Brianna Wu and the human cost of Gamergate: ‘Every woman I know in the industry is scared’”, *The Guardian*, 18 October. Accessed from <http://www.theguardian.com/technology/2014/oct/17/brianna-wu-gamergate-human-cost> on 8 November 2018.
- [45] Stuart, Keith (2014b), “Zoe Quinn: ‘All Gamergate has done is ruin people’s lives’”, *The Guardian*, 4 December. Accessed from <http://www.theguardian.com/technology/2014/dec/03/zoe-quinn-gamergate-interview> on 8 November 2018.

- [46] Marcetic, Branko (2014), “#Gamergate is really about terrorism: Why Bill Maher should be vilifying the gaming community, too”, Salon, 24 October. Accessed from http://www.salon.com/2014/10/23/gamergate_is_really_about_terrorism_why_bill_maher_should_be_vilifying_the_gaming_community_too/ on 8 November 2018.
- [47] Colvin, Mark, and Mark, David (2012), “TV presenter in hospital after vicious Twitter attacks”, PM, 30 August. Accessed from <http://www.abc.net.au/pm/content/2012/s3579714.htm> on 8 November 2018.
- [48] Brown, Tara (2012), “Charlotte’s hell”, 60 Minutes, 31 August. Accessed from <http://sixtyminutes.ninemsn.com.au/article.aspx?id=8525498> on 8 November 2018.
- [49] Pariser, Eli (2011), *The Filter Bubble: What the Internet Is Hiding from You*. London: Viking.
- [50] Valenti, Jessica (2016), “Insults and rape threats: Writers shouldn’t have to deal with this”, The Guardian, 15 April. Accessed from <http://www.theguardian.com/commentisfree/2016/apr/14/insults-rape-threats-writers-online-harassment> on 8 November 2018.
- [51] “Australian media still a Blokesworld in 2016” (2016), Media, Entertainment and Arts Alliance, 6 March. Accessed from <https://www.meaa.org/news/australianmedia-still-a-blokesworld-in-2016/> on 8 November 2018.
- [52] Boggioni, Tom (2016), “Prominent Feminist Writer Drops off Social Media after Rape Threat against Her 5-Year-Old Daughter”, Raw Story, 27 July. Accessed from <http://www.rawstory.com/2016/07/prominent-feminist-writer-drops-off-social-media-after-rape-threats-against-her-5-year-olddaughter/> on 1 March 2017.
- [53] West, Lindy (2017), “I’ve Left Twitter. It is Unusable for Anyone but Trolls, Robots and Dictators”, The Guardian, 4 January. Accessed from <https://www.theguardian.com/commentisfree/2017/jan/03/ive-left-twitter-unusable-anyone-but-trolls-robots-dictators-lindy-west> on 27 October 2017.
- [54] West, Lindy (2015), “What Happened When I Confronted My Cruellest Troll”, The Guardian, 3 February. Accessed from <http://www.theguardian.com/society/2015/feb/02/what-happenedconfronted-cruellest-troll-lindy-west> on 23 February 2017.

- [55] Citron, Danielle Keats (2014b), “‘Revenge porn’ should be a crime in U.S.,” CNN, 16 January. Accessed from <http://edition.cnn.com/2013/08/29/opinion/citron-revenge-porn/> on 16 January 2016.
- [56] Citron, Danielle Keats (2014c), “How cyber mobs and trolls have ruined the internet – and destroyed lives”, Newsweek, 19 September. Accessed from <http://www.newsweek.com/internet-and-golden-age-bully-271800> on 14 January 2016.
- [57] “Amnesty International, Australia: Poll reveals alarming impact of online abuse against women” (2018), Amnesty International Australia, 7 February. Accessed from <https://www.amnesty.org.au/australia-poll-reveals-alarming-impact-online-abuse-women/> on 8 November 2018.
- [58] “Australian Human Rights Commission Human Rights and Technology Issues Paper July 2018” (2018), Australian Human Rights Commission, 24 July. Accessed from <https://tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf> on 9 November 2018.
- [59] Junger, Marianne (2018), “On the prevalence of cybercrime across Europe”, BioMed Central, 22 August. Accessed from <http://blogs.biomedcentral.com/on-health/2018/08/22/on-the-prevalence-of-cybercrime-across-europe/> on 9 November 2018.
- [60] Farrell, Graham, Tilley, Nick, and Tseloni, Andromachi (2014), “Why the Crime Drop?”, *Crime and Justice*, September: 43 (1): pp. 421-490.
- [61] Kranenbarg, Marleen Weulen, Holt, Thomas J., and van Gelder, Jean-Louis (2017), “Offending and Victimization in the Digital Age: Comparing Correlates of Cybercrime and Traditional Offending-Only, Victimization-Only and the Victimization-Offending Overlap”, *Deviant Behavior*, December. DOI: 10.1080/01639625.2017.1411030.
- [62] Jane, Emma A. and Vincent, Nicole A (2016) “Random Rape Threat Generator: Radio Edit”. Accessed from <https://www.rapeglish.com/> on 4 October 2018.
- [63] Jane, Emma A. and Vincent, Nicole A (2016) “Random Rape Threat Generator: Extended Remix”. Accessed from https://rapethreatgenerator.com/?page_id=917 on 4 October 2018.

- [64] Jane, Emma A. and Vincent, Nicole A (18 July 2017) “Women online are getting used to cyber hate. They need to get used to reporting it”, Sydney Morning Herald. Accessed from <http://www.smh.com.au/lifestyle/news-and-views/opinion/women-online-are-getting-used-to-cyber-hate-they-need-to-get-used-to-reporting-it-20170717-gxctr8.html> on 19 July 2017.
- [65] Vincent, Nicole A and Jane, Emma A. (18 July 2017) “A crime is a crime, even if it's online — here are six ways to stop cyberhate”, Australian Broadcasting Corporation. Accessed from <http://www.abc.net.au/news/2017-07-18/six-ways-to-stop-cyberhate/8721184> on 19 July 2017.
- [66] Vitis, Laura and Segrave, Marie (2017), “Introduction”, in Segrave, Marie, and Vitis, Laura (eds.), *Gender, Technology and Violence*. London and New York: Routledge.
- [67] Ostini, Jenny, and Hopkins, Susan (2015) “Online harassment is a form of violence”, *The Conversation*, 8 April. Accessed from <https://theconversation.com/online-harassment-is-a-form-of-violence-38846> on 11 January 2016.
- [68] “Inside Facebook” (2018), *Four Corners*, ABC-TV, aired 6 August.
- [69] Schipp, Debbie (2018), “Toxic content their ‘crack cocaine’: Facebook’s disturbing moderator secrets”, *news.com.au*, 7 August. Accessed from <https://www.news.com.au/entertainment/tv/current-affairs/toxic-content-their-crack-cocaine-facebooks-disturbing-moderator-secrets/news-story/e03358922d893e49286fe514b11fe504> on 10 November 2018.
- [70] Shaw, Francis (2017). *Information, Communication & Society*, 20(12): 1783-1785.
- [71] “SlutWalk” (n.d.), Facebook. Accessed from https://www.facebook.com/pg/SlutWalk/about/?ref=page_internal on 11 November 2018.
- [72] “Our Mission” (n.d), Women’s March. Accessed from <https://www.womensmarch.com/mission/> on 11 November 2018.
- [73] “Web Index: Report 2014–15” (n.d.), Web Index. Accessed from https://thewebindex.org/wp-content/uploads/2014/12/Web_Index_24pp_November2014.pdf on 11 November 2018.

[74] Vincent, Nicole A (2017), “Victims of cybercrime: Definitions and challenges”, in Martellozzo, Elena and Jane, Emma J. (eds.), *Cybercrime and its Victims*. Oxon: Routledge, pp. 27-42.